

普通高等教育"十一五"国家级规划教材

教育部"高等学校教学质量与教学改革工程"立项项目

张兴会 等编著

数据仓库与数据挖掘工程实例

计算机科学与技术专业实践系列教材

清华大学出版社

计算机科学与技术专业实践系列教材

数据仓库与数据挖掘工程实例

张兴会 等 编著

清华大学出版社
北 京

内 容 简 介

数据仓库与数据挖掘是与计算机、信息类等相关专业的核心课程。本书采用提出问题、分析问题、解决问题的思路，通过工程实例介绍了 SQL Server 2005 和 Weka 软件的使用方法以及联机分析处理技术、关联规则方法、决策树方法、贝叶斯方法、人工神经网络方法、聚类分析方法、线性回归方法等数据仓库与数据挖掘技术。

本书结构严谨，条理清晰，语言浅显易懂，循序渐进地表达了知识内容；坚持理论与实际相结合，知识理论与具体实现方法相结合，使技术实现具体化、生动化、可操作化；工程实例的实现过程建立在 SQL Server 2005 和 Weka 软件的基础上，以帮助读者在学习后达到学以致用的效果。本书可以和《数据仓库与数据挖掘技术》教材配合使用，旨在帮助读者在学习数据仓库与数据挖掘理论知识的基础上，通过学习工程实例分析，较好地掌握数据挖掘与数据仓库构建模型的操作过程，进一步提高对信息管理和利用能力。

本书可以作为计算机、信息类等专业本科生数据挖掘课程的教材，也可以作为其他专业技术人员的自学参考书。

图书在版编目(CIP)数据

数据仓库与数据挖掘工程实例/张兴会等编著. --北京：清华大学出版社，2014
计算机科学与技术专业实践系列教材
ISBN 978-7-302-35541-0

Ⅰ. ①数…　Ⅱ. ①张…　Ⅲ. ①数据库系统－高等学校－教材 ②数据采集－高等学校－教材 Ⅳ. ①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字(2014)第 034937 号

责任编辑：汪汉友　赵晓宁
封面设计：
责任校对：李建庄
责任印制：

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：8.5　　**字　　数**：212 千字
版　　次：2014 年 9 月第 1 版　　**印　　次**：2014 年 9 月第 1 次印刷
印　　数：1～000
定　　价：　.00 元

产品编号：034552-01

前　　言

数据挖掘技术在科学研究和日常生活中具有广泛的应用，被列为 21 世纪最具潜力的应用技术之一。现在数据挖掘技术已经成为信息系统、应用数学等专业学生的必修教学内容。

为此，本书在编写时力求突出以下特色：

(1) 引入数据挖掘研究的热点问题以及最新研究成果，保证教材的先进性。

(2) 强化目标驱动观点，使读者学习有的放矢。

(3) 每章后面都详细讲解了在 SQL Server 2005 或 Weka 环境下相关理论的具体实现技术，使得读者可以理论联系实际，培养解决实际问题的能力。

(4) 在文字表达方面争取语言更通俗、易懂、易读。

本书具体内容如下：

实例 1～实例 10，详细介绍了基于联机分析处理技术、关联规则方法、决策树方法、贝叶斯方法、人工神经网络方法、聚类分析方法、线性回归方法等方法的 10 个工程实例的具体实现。

附录 A 和附录 B，分别介绍了 SQL Server 2005 和 Weka 软件的任务描述和实现方法。

本书的案例来源于不同的专业领域和最新的工程实践，新颖独特，具有代表性和很强的实际借鉴价值。读者通过学习，可以了解和掌握数据挖掘技术的理论和算法，熟悉在各个领域应用的流程和分析方法，从而为以后的数据分析工作夯实基础。

为了能更好地将工程实例与相关理论知识相结合，将基本概念与具体的方法、工具相结合，达到学以致用的效果，读者可参考笔者所编著的《数据仓库与数据挖掘技术》进行学习。

本书由张兴会统稿，王明春、郑晓艳、刘玲、刘新钰参加了本书的编写、图表绘制、模型构建、软件调试等工作。在本书编写过程中，安淑芝教授提出了宝贵的修改意见。另外，本书还参阅和引用了许多专家和学者的文献资料，在此表示衷心的感谢。

由于笔者水平和能力有限，新技术的发展和更新较快，书中难免有不妥之处，欢迎读者批评指正。笔者邮箱为 xhzhang@tute.edu.cn。

编　者

2014 年 8 月

目　　录

实例 1　基于联机分析处理技术的税务审计分析

1.1　任 务 描 述

需要对某市国税局的延期纳税审批情况进行审计，资料来源于某市国税局延期纳税数据库，此数据库中共有三个数据表。

（1）延期纳税批件表：在此表中共有 1568 条记录，记录着税务局批准企业纳税的基本信息，例如征收项目种类、税款所属期、税额、纳税人名称等。

（2）税务机关代码表：记录该市所属各区县的税务机关代码及名称。

（3）征收项目代码表：记录各征收项目税种的代码及名称。

在审计时，面临诸多的困难，如时间跨度大（从 2002 年 1 月至 2004 年 2 月）、所属区县多、审批金额大等。对税务局审批延期纳税的合法合规性分析离不开对纳税企业的延伸。如何在这些浩如烟海的电子资料中找到需要的信息是这次审计的核心问题。

本例将通过分析数据库中国税局给企业批准延期纳税的大量数据，简要介绍审计过程中如何应用多维数据分析工具在统揽全局、把握总体的基础上对大量的电子数据进行筛选、分析，快速找出审计重点，准确定位延伸分析的对象。通过对某市国税局的深入了解，在审计时需要掌握以下情况：2002—2004 年全市共审核批准了多少延期纳税税款？哪年审批的金额比较大？审计的都有什么税种？各税种占的比例有多大？各个区县分别审批了多少税款？哪个区县审批的金额较多？审批时间集中在什么时候？审计人员应从哪里进行突破？

在这个案例中，通过建立多维数据集，在把握总体、统揽全局的基础上观察趋势，选择重点；最后进行有针对性的延伸取证，顺利完成了整个审计过程。

1.2　技 术 原 理

1.2.1　联机分析处理的定义

联机分析处理委员会对联机分析处理（OLAP）的定义为：使分析、管理或执行人员能够从多种角度对从原始数据中转化出来、能够真正为用户所理解、并真实反映企业维特性的信息进行快速、一致、交互地存取，从而获得对数据更深入了解的一类软件技术。

OLAP 的基本多维分析操作有钻取（Drill-up 和 Drill-down）、切片（Slice）和切块（Dice）以及旋转（Pivot）等。

1.2.2　联机分析处理的一些具体操作

1. 钻取

钻取是改变维的层次，变换分析的粒度。它包括向下钻取（Drill-down）和向上钻取（Drill-up）。向上钻取也称为上卷（Roll-up），是在某一维上将低层次的细节数据概括到高层次的汇总数据，或减少维数；而 Drill-down 则相反，它从汇总数据深入到细节数据进行观察

或增加新维。例如,图 1-1 所示的数据立方体经过沿着分行维的概念层次上卷,由分行上升到城市,得到如图 1-2 所示的立方体;图 1-1 中的数据立方体经过沿时间维下钻,由年度下降到季度,得到如图 1-3 所示的数据立方体。

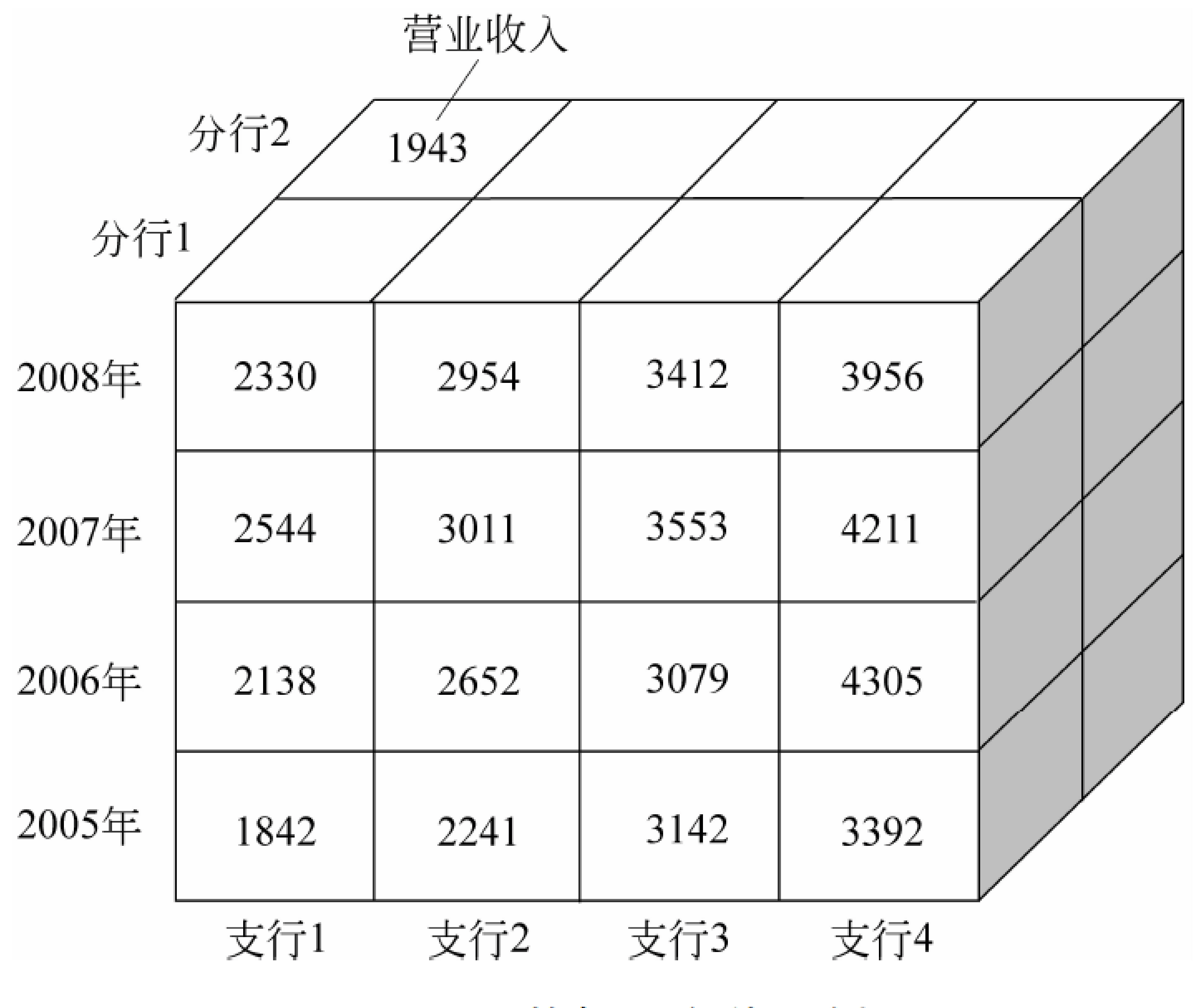


图 1-1　数据立方体示例

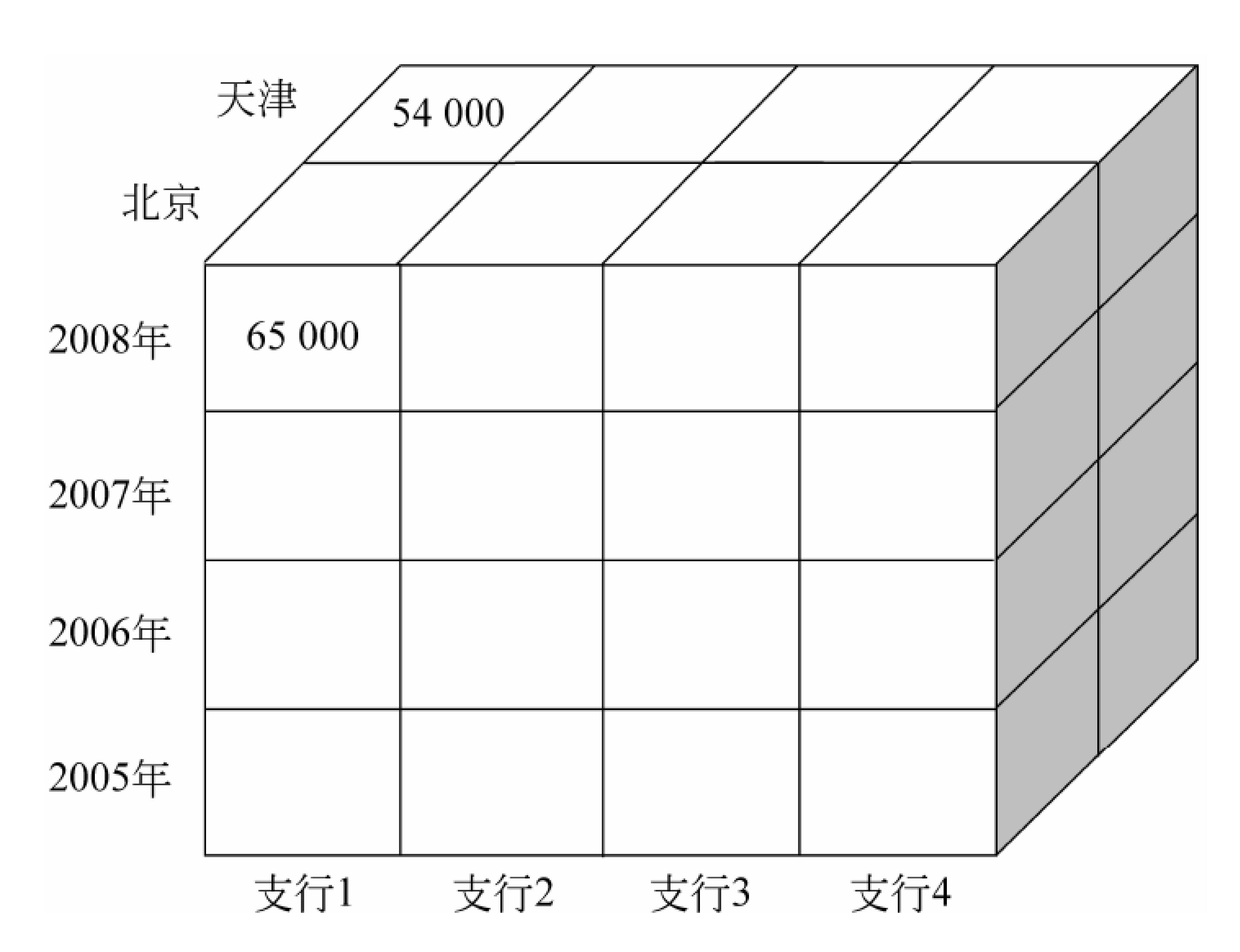


图 1-2　向上钻取后得到的数据立方体

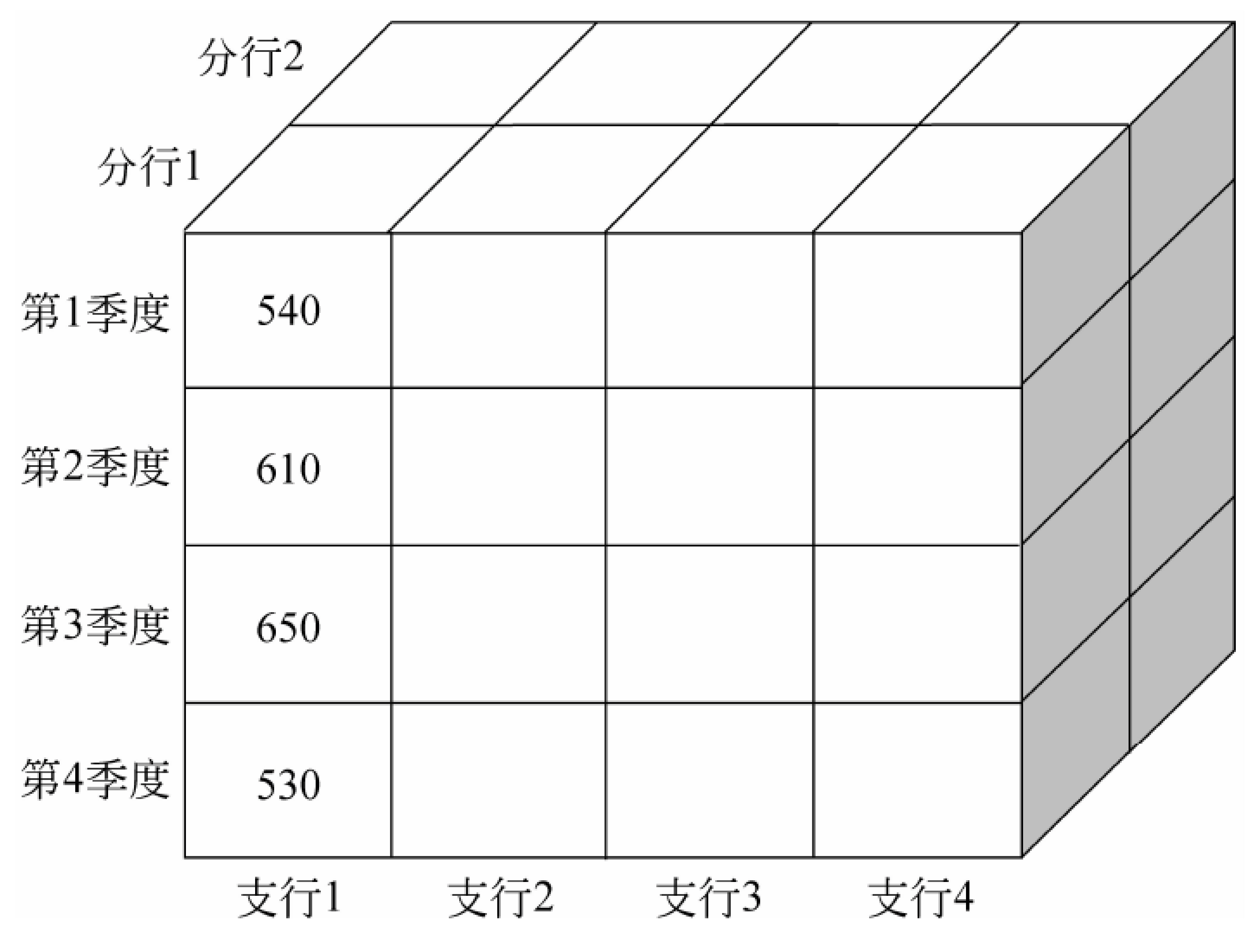


图 1-3　向下钻取后得到的数据立方体

2. 切片和切块

切片：在给定数据立方体的一个维上进行选择操作就是切片，切片的结果是得到一个二维平面数据。例如，对图 1-1 中数据立方体，使用条件：

```
“银行分行=‘分行 1’”
```

进行选择，就相当于在原来的立方体中切出一片，结果如图 1-4 所示。

切块：在给定数据立方体的两个或多个维上进行选择操作就是切块，切块的结果得到一个子立方体。例如，对图 1-1 所示数据立方体，使用条件：

```
(银行分行=“分行 1”or“分行 2”)
And (时间=“2007 年”or“2008 年”)
And (银行支行=“支行 1”or“支行 2”)
```

进行选择，就相当于在原立方体中切出一小块，结果如图 1-5 所示。

	支行1	支行2	支行3	支行4
2008年	2330	2954	3412	3956
2007年	2544	3011	3553	4211
2006年	2138	2652	3079	4305
2005年	1842	2241	3142	3392

图 1-4　切片示例

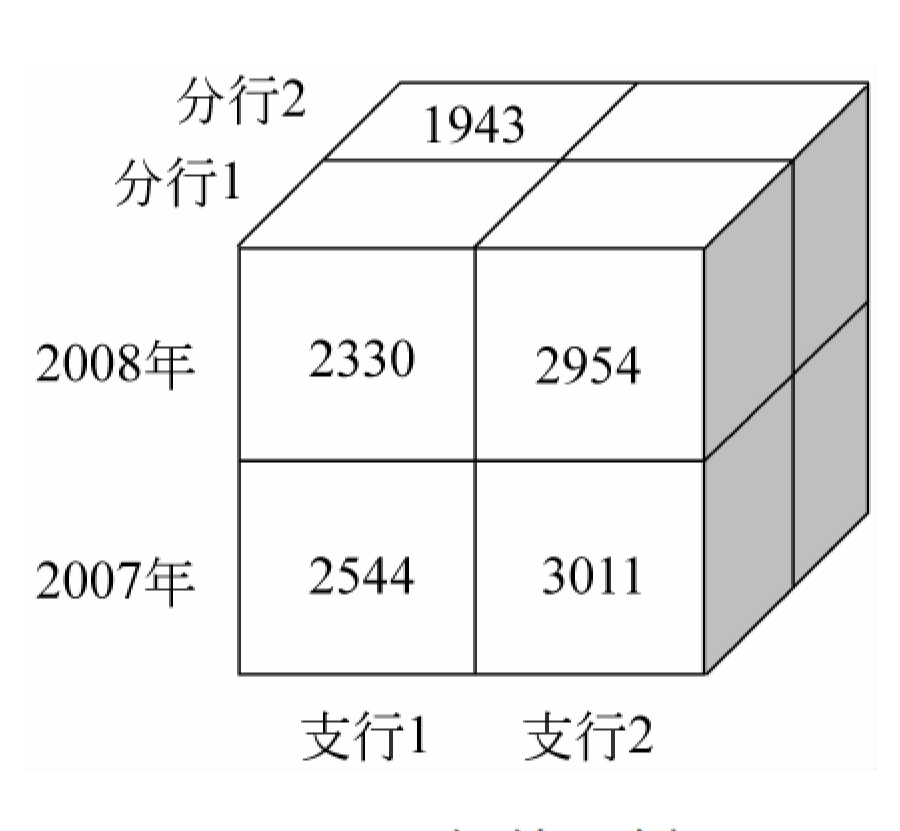


图 1-5　切块示例

3. 旋转

旋转是变换维的方向，即在表格中重新安排维的放置（如行列互换）。图 1-6 所示是图 1-1 中立方体通过旋转横纵坐标所得的立方体。

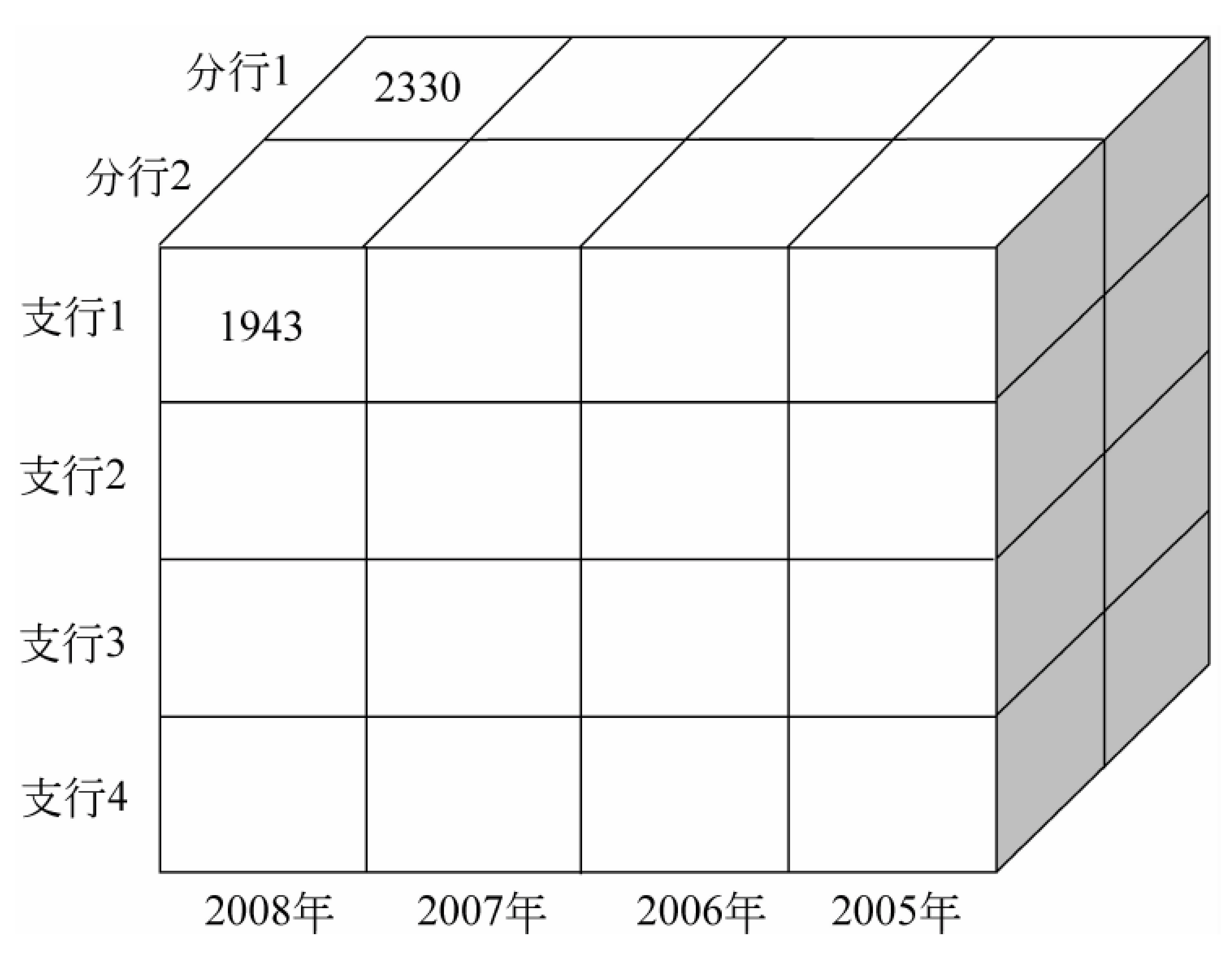


图 1-6　旋转后得到的数据立方体

1.3 具体实现

1.3.1 建立数据库

建立数据库的步骤如下：

(1) 依次执行“开始”→“程序”→ Microsoft SQL Server 2005 → SQL Server Management Studio 命令，如图 1-7 所示，打开 SQL Server 2005 数据库管理器。

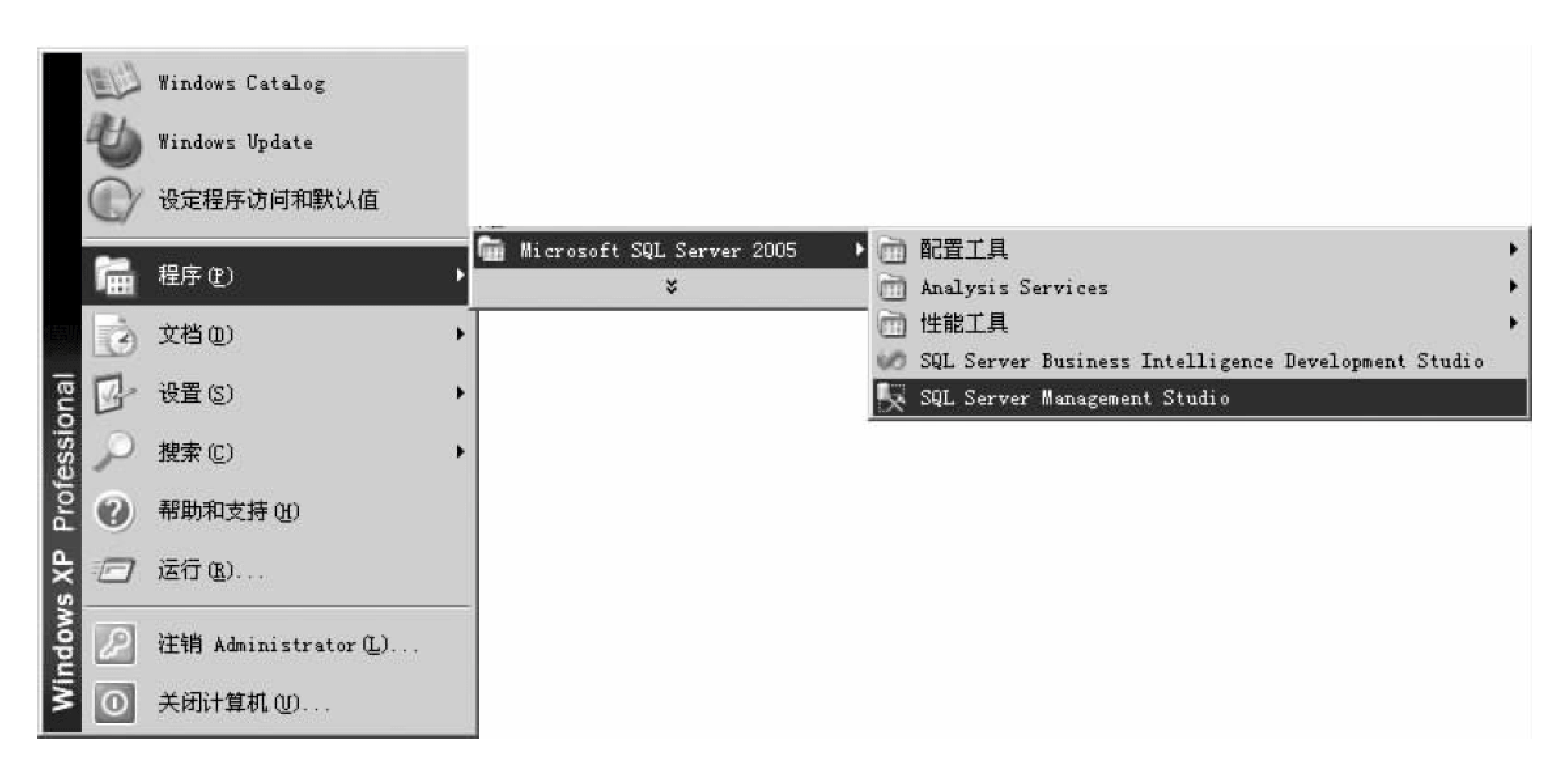


图 1-7 打开数据库管理器

(2) 在弹出“连接服务器”对话框中选择安装 SQL Server 2005 时所建立的命名实例名，在身份验证中选择“Windows 身份验证”项，单击“连接”按钮，如图 1-8 所示。



图 1-8 连接服务器

(3) 进入“对象资源管理器”界面后，在左侧树形结构中找到“数据库”文件夹，右击，在弹出的快捷菜单中选择“新建数据库”命令，如图 1-9 所示。

(4) 在弹出的“新建数据库”对话框的“数据库名称”文本框中填写“延期纳税”，单击“确定”按钮，如图 1-10 所示。

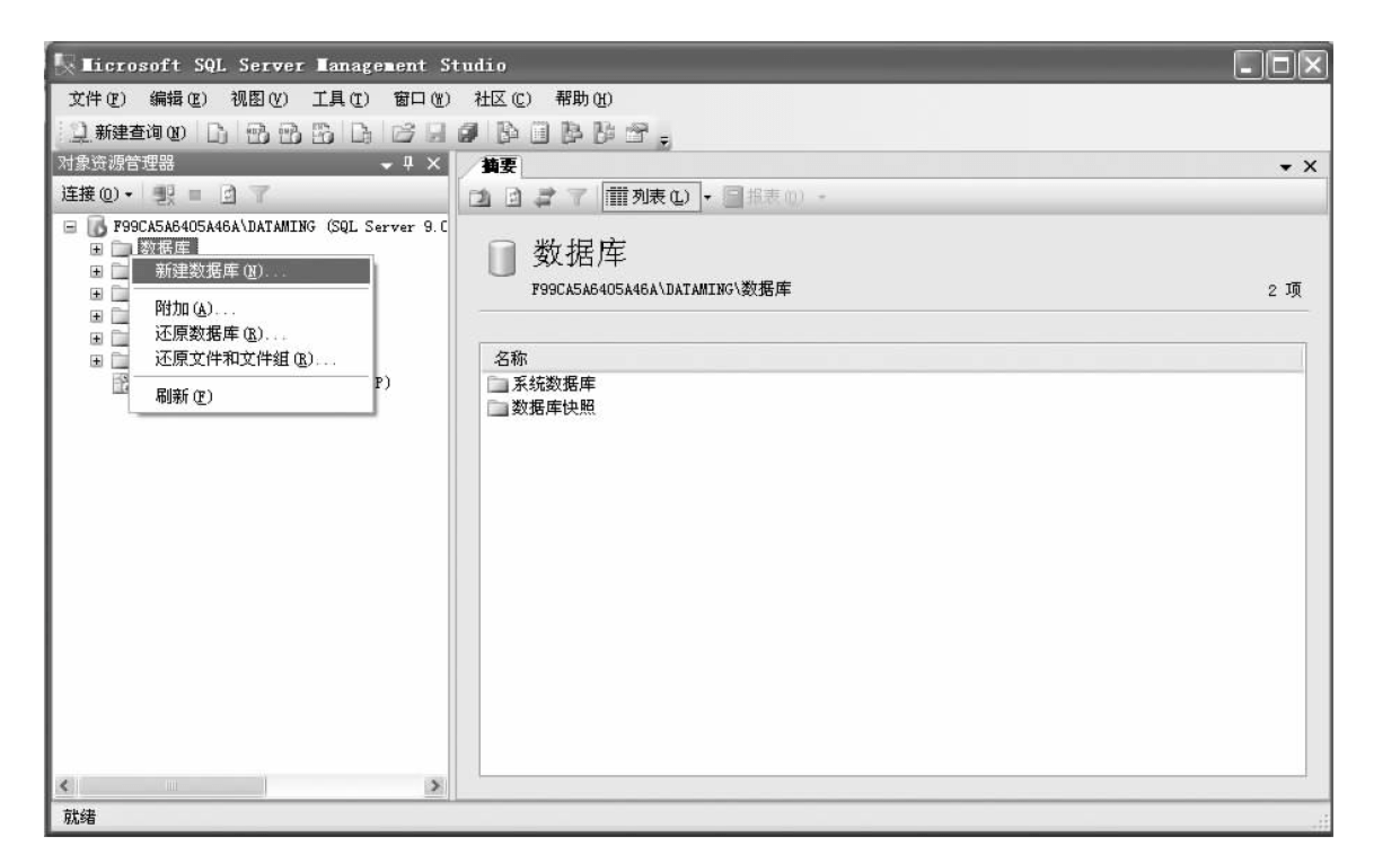


图 1-9　进入对象资源管理器

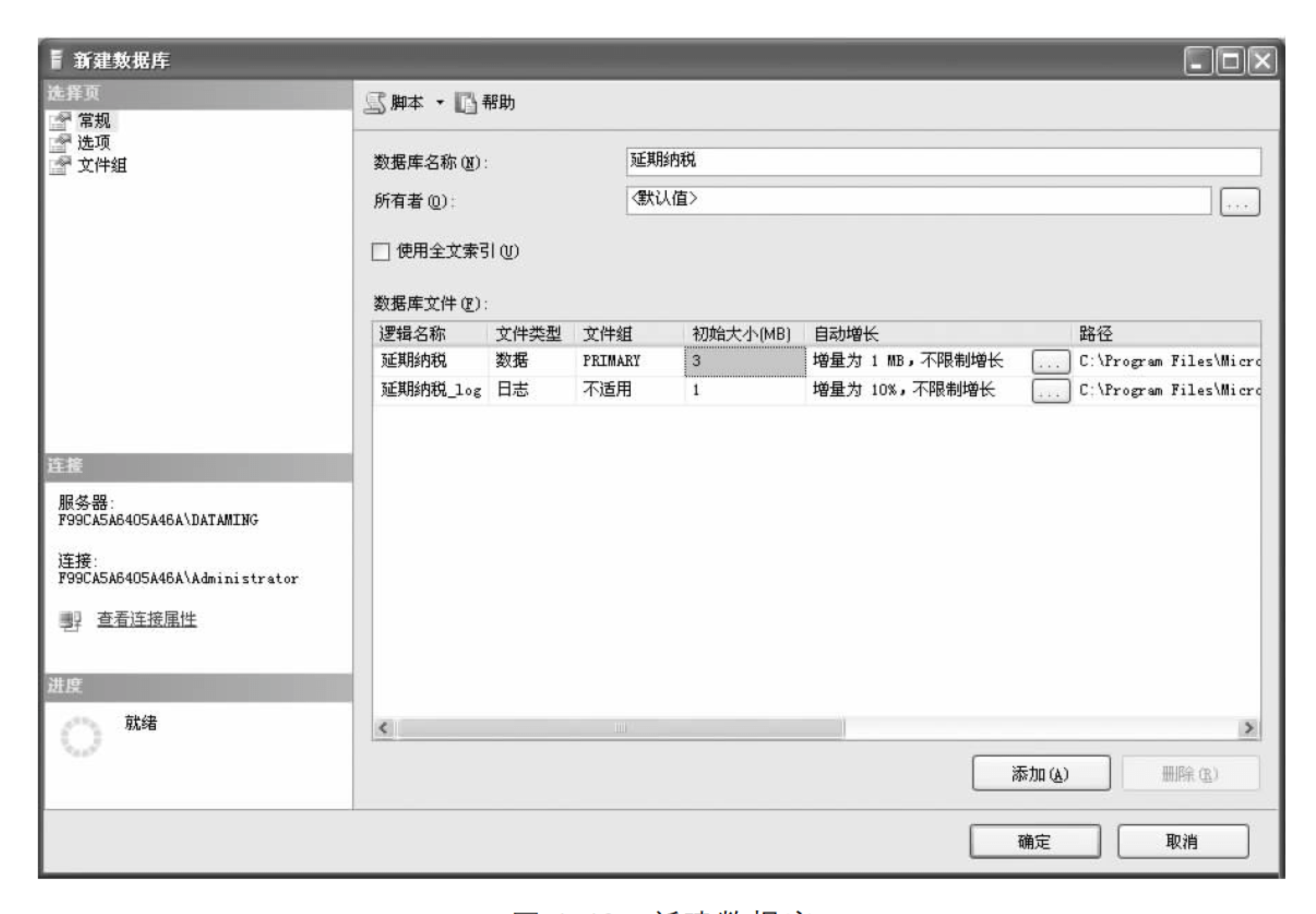


图 1-10　新建数据库

（5）回到“对象资源管理器”界面，在左侧树形结构中找到新建立的数据库“延期纳税”项，右击“延期纳税”数据库，在弹出的快捷菜单中选择“任务”→“导入数据”命令，如图 1-11 所示。

（6）打开“SQL Server 导入和导出向导”对话框，如图 1-12 所示。

（7）单击“下一步”按钮。在“数据源”下拉列表中选择 Microsoft Access 项，如图 1-13 所示。

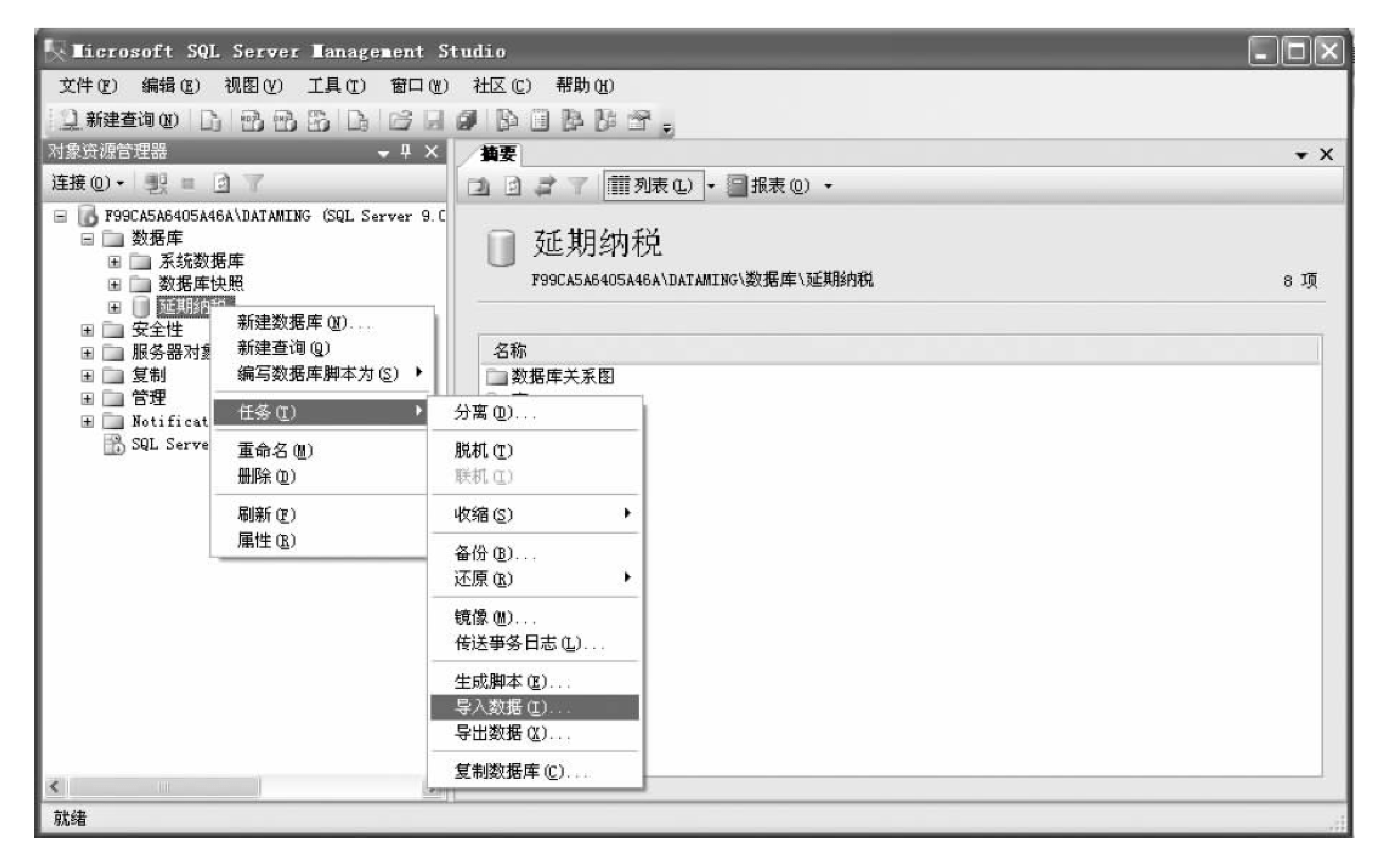


图 1-11 选择“导入数据”选项



图 1-12 打开导入和导出向导

SQL Server 导入和导出向导
选择数据源
选择要从中复制数据的源。
数据源(D): Microsoft Access
Microsoft Access
Microsoft Excel
Microsoft OLE DB Provider for Analysis Services 9.0
Microsoft OLE DB Provider For Data Mining Services
Microsoft OLE DB Provider for Internet Publishing
Microsoft OLE DB Provider for OLAP Services 8.0
Microsoft OLE DB Provider for Oracle
Microsoft OLE DB Provider for SQL Server
若要进行连接，请选择数据库并扌
文件名(I):
用户名(U):
密码(P):
高级(A)...
帮助(H)
< 上一步(B)
下一步(N) >
完成(F) >>|
取消

图 1-13 选择数据源类型

(8) 单击“下一步”按钮。选择需要导入的数据，单击“打开”按钮，如图 1-14 所示。



图 1-14　选择导入数据

(9) 在弹出的“选择数据源”页面中，单击“下一步”按钮，如图 1-15 所示。



图 1-15　确定导入数据

(10) 在弹出的“选择目标”页面中，单击“下一步”按钮，如图 1-16 所示。

(11) 在弹出的“指定表复制或查询”页面中选择“复制一个或多个表或视图的数据”单选按钮并单击“下一步”按钮，如图 1-17 所示。

(12) 在弹出的“选择源表和源视图”页面中，单击“全选”按钮，如图 1-18 所示。所有需要导入的数据表全部被选中，单击“下一步”按钮。

(13) 单击“预览”对导入数据进行预览，并单击“确定”按钮，如图 1-19 所示。

图 1-16 选择目标

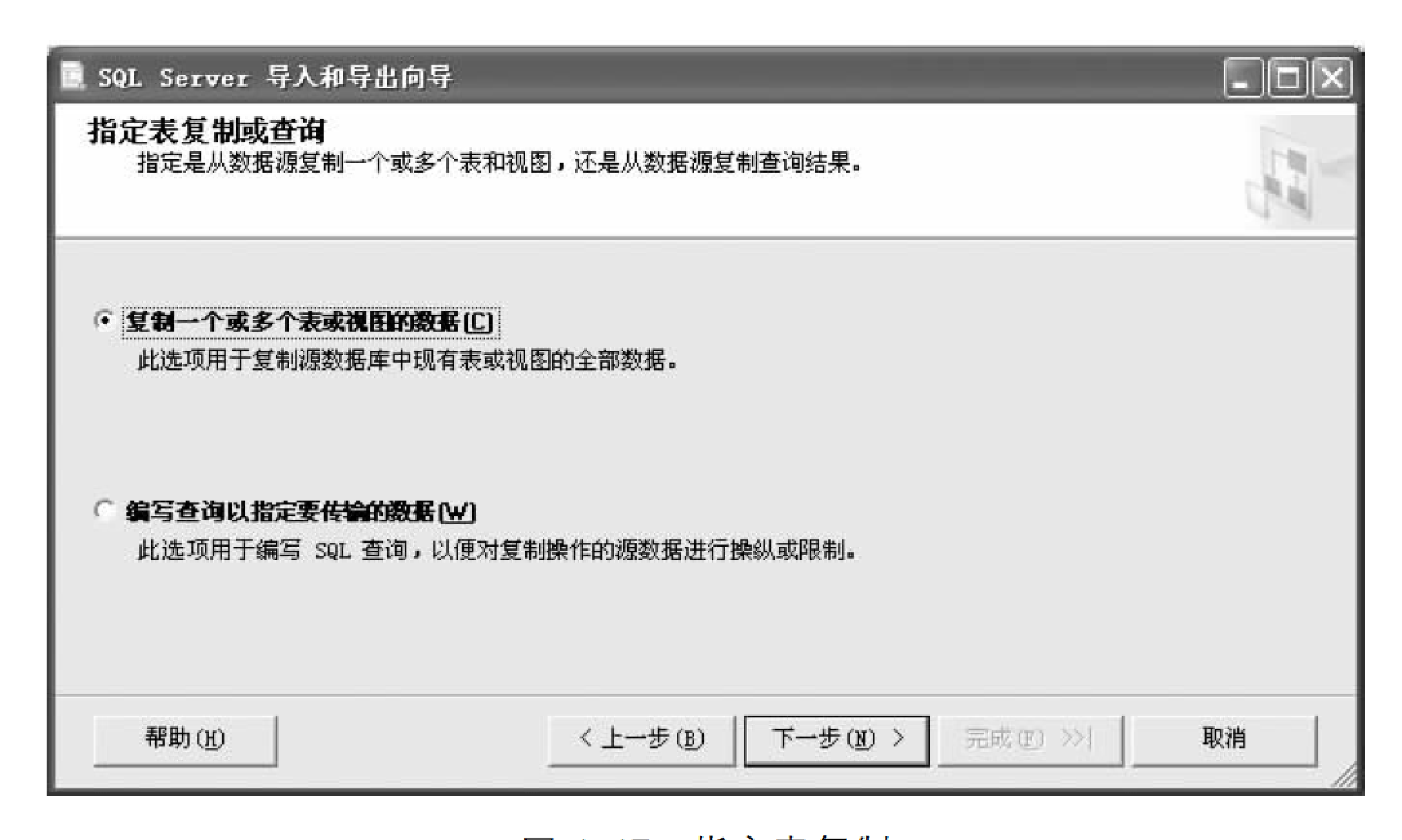
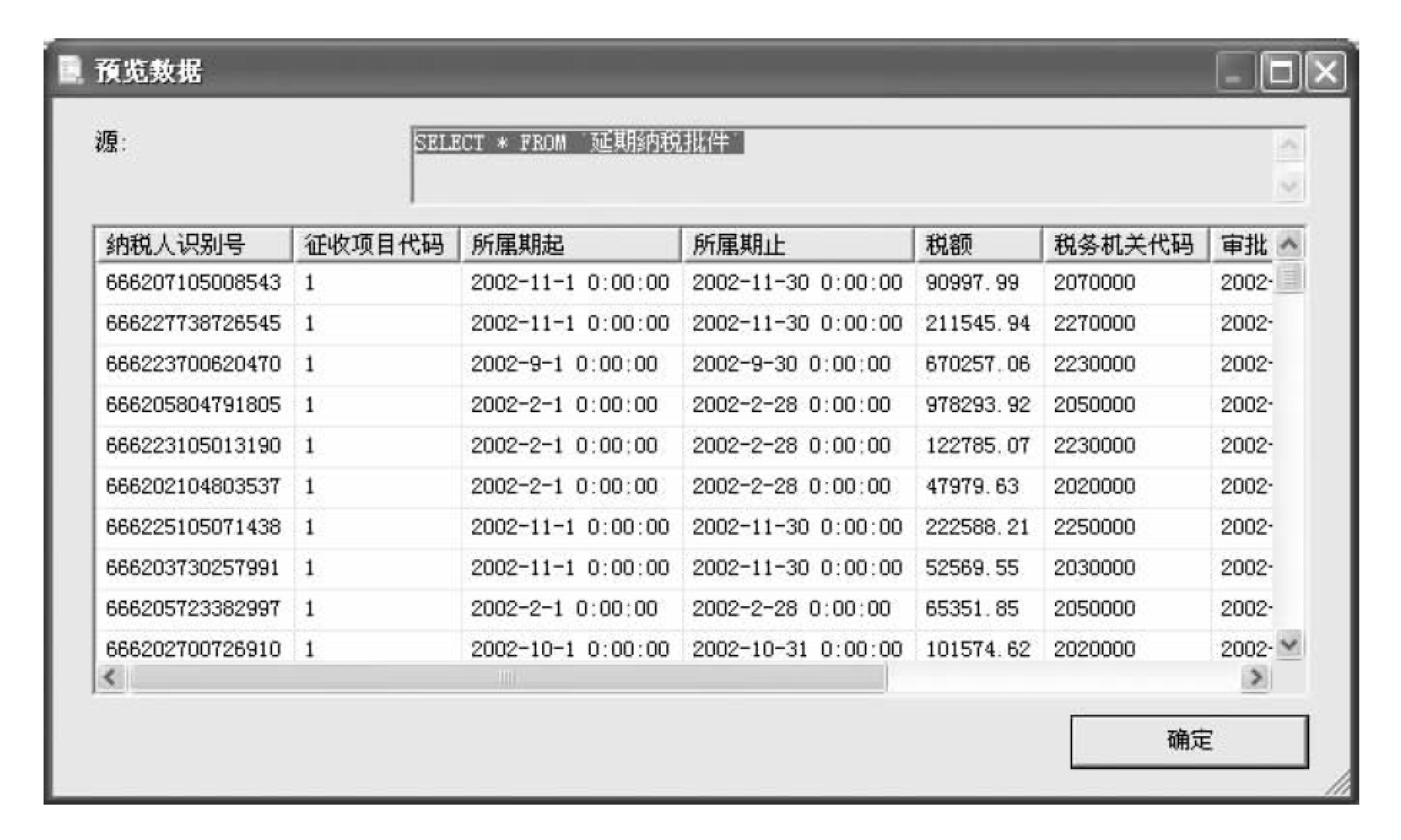


图 1-17 指定表复制



图 1-18 选择源表



纳税人识别号	征收项目代码	所属期起	所属期止	税额	税务机关代码	审批
666207105008543	1	2002-11-1 0:00:00	2002-11-30 0:00:00	90997.99	2070000	2002-
666227738726545	1	2002-11-1 0:00:00	2002-11-30 0:00:00	211545.94	2270000	2002-
666223700620470	1	2002-9-1 0:00:00	2002-9-30 0:00:00	670257.06	2230000	2002-
666205804791805	1	2002-2-1 0:00:00	2002-2-28 0:00:00	978293.92	2050000	2002-
666223105013190	1	2002-2-1 0:00:00	2002-2-28 0:00:00	122785.07	2230000	2002-
666202104803537	1	2002-2-1 0:00:00	2002-2-28 0:00:00	47979.63	2020000	2002-
666225105071438	1	2002-11-1 0:00:00	2002-11-30 0:00:00	222588.21	2250000	2002-
666203730257991	1	2002-11-1 0:00:00	2002-11-30 0:00:00	52569.55	2030000	2002-
666205723382997	1	2002-2-1 0:00:00	2002-2-28 0:00:00	65351.85	2050000	2002-
666202700726910	1	2002-10-1 0:00:00	2002-10-31 0:00:00	101574.62	2020000	2002-

图 1-19　预览数据

(14) 在“保存并执行包”页面中,单击“下一步”按钮,如图 1-20 所示。



图 1-20　保存并执行包

(15) 在弹出的“完成该向导”页面中,单击 “完成”按钮,如图 1-21 所示。

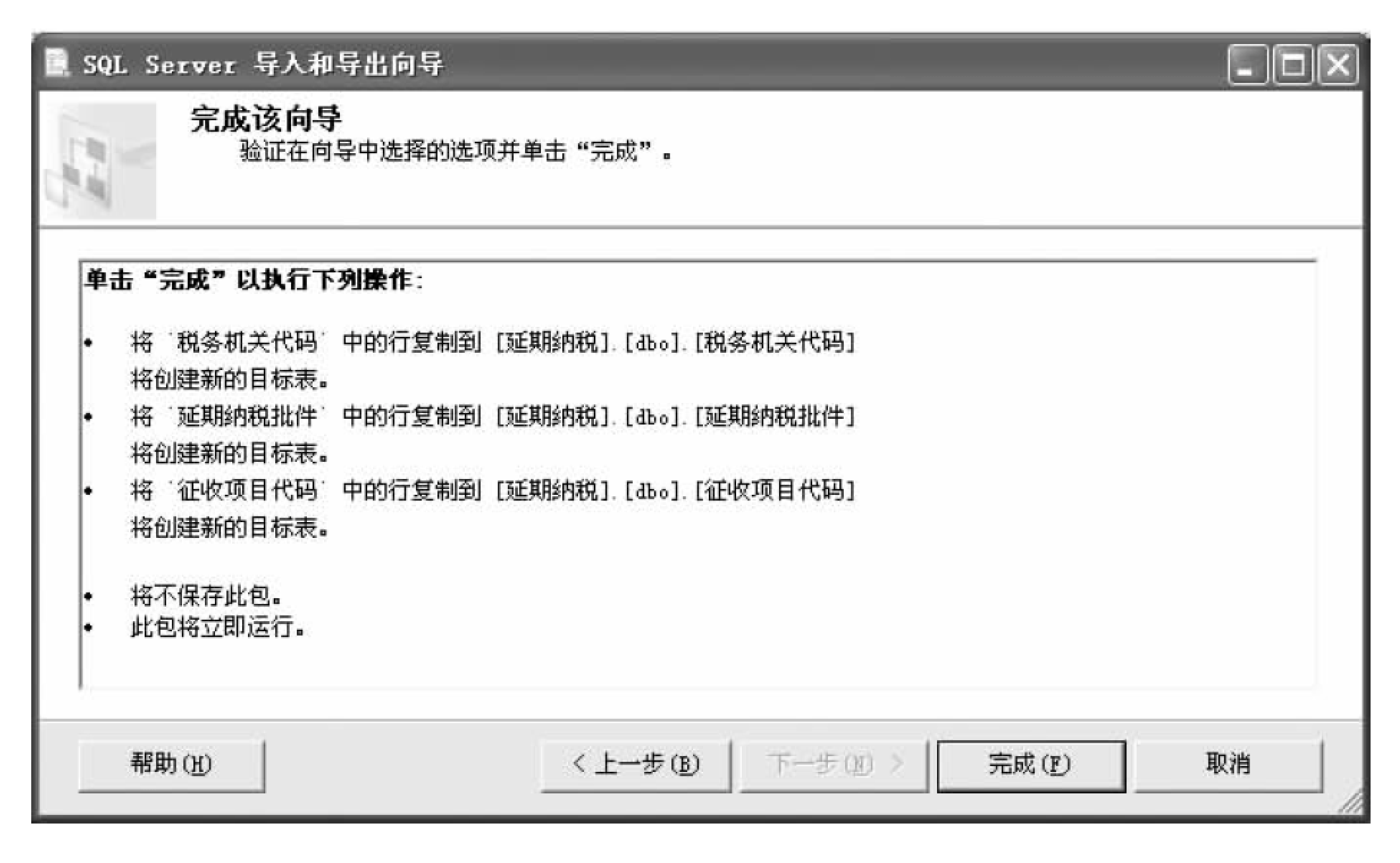


图 1-21　完成导入和导出向导

（16）在弹出的“执行成功”对话框中，单击“关闭”按钮，完成数据库的建立，如图 1-22 所示。



图 1-22　完成数据库建立

1.3.2　新建数据源

新建数据源的操作步骤如下：

（1）选择“开始”→“程序”→Microsoft SQL Server 2005→SQL Server Business Intelligence Development Studio 进入 Business Intelligence Development Studio(BIDS)，如图 1-23 所示。

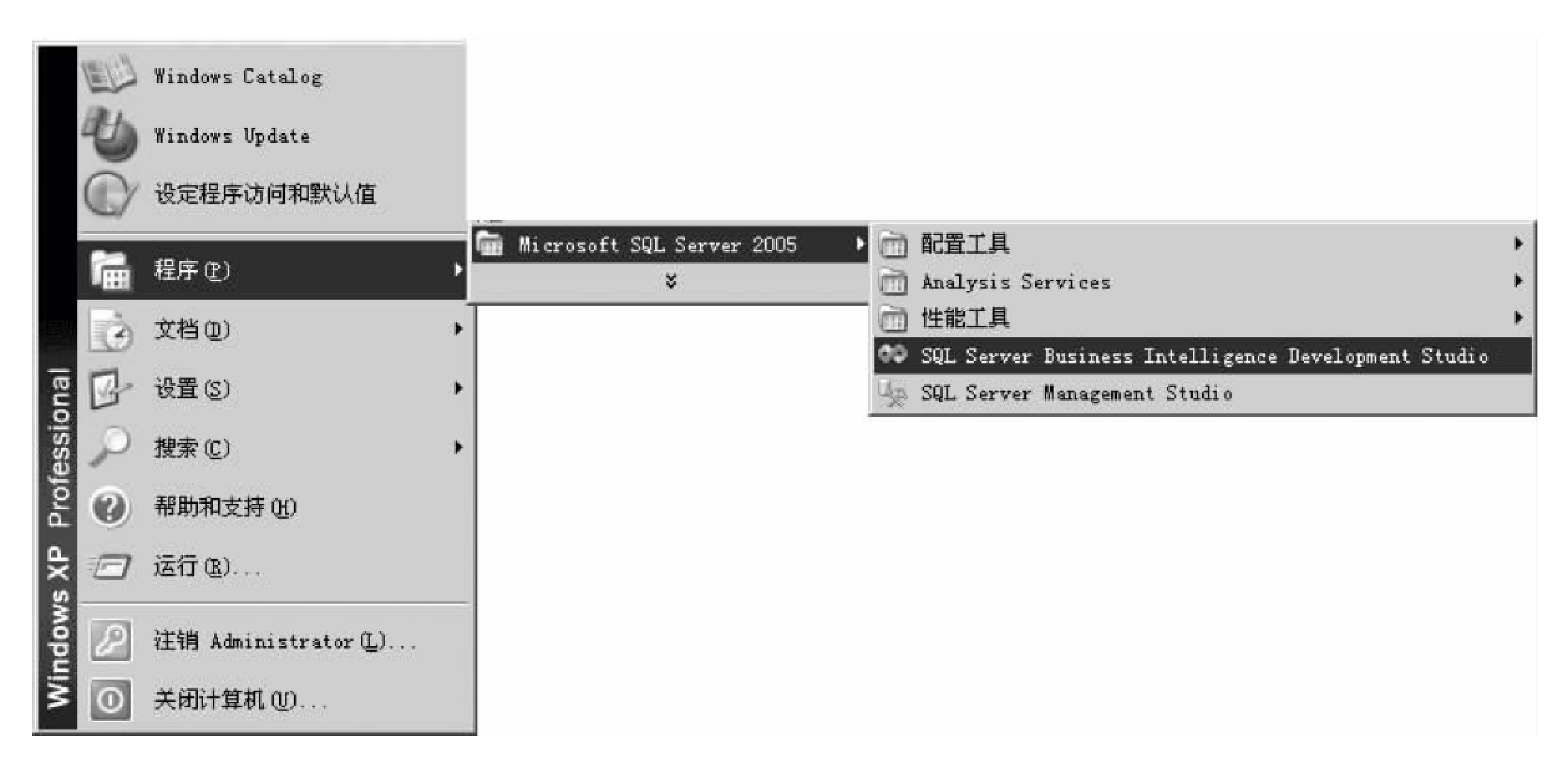


图 1-23　进入 BIOS

（2）选择“文件”→“新建”→“项目”命令，如图 1-24 所示。

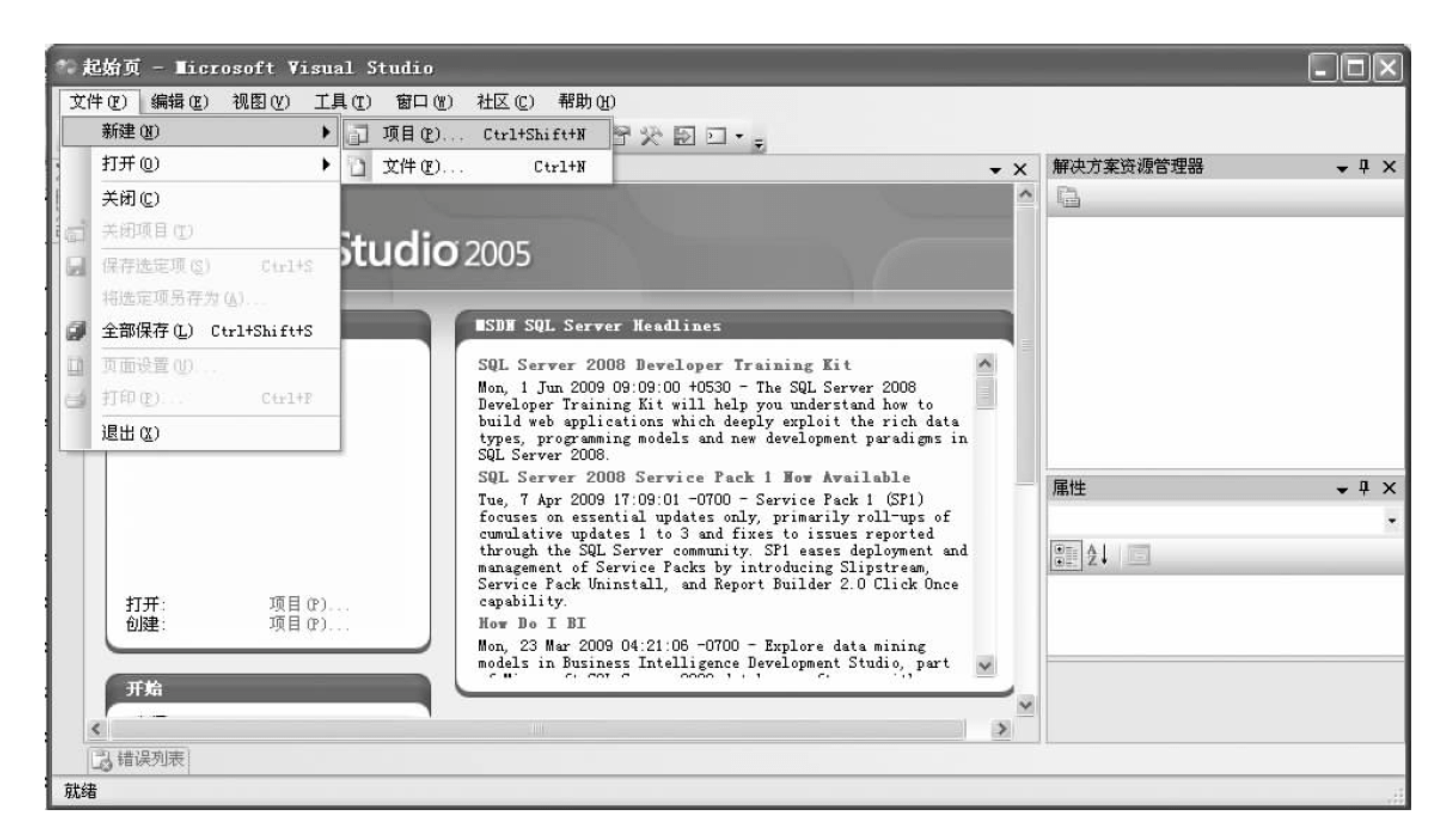


图 1-24　打开“新建项目”选项

(3) 在“新建项目”对话框中选择项目类型“商业智能下的 Analysis Services 项目”，项目名称为“延期纳税”，单击“确定”按钮，创建数据挖掘项目，如图 1-25 所示。



图 1-25　创建数据挖掘项目

(4) 在“延期纳税”的解决方案资源管理器中，右击“数据源”项，在弹出的快捷菜单中选择“新建数据源”命令，如图 1-26 所示。

(5) 在弹出的“数据源向导”对话框中，单击“下一步”按钮，如图 1-27 所示。

(6) 在弹出的“选择如何定义连接”页面中，单击“新建”按钮，如图 1-28 所示。

(7) 在弹出的“连接管理器”对话框中，设置服务器名为前面建立的命名空间名，选择“使用 Windows 身份验证”单选按钮，数据库选择之前创建的“延期纳税”，单击“测试连接”按钮，如图 1-29 所示。



图 1-26 打开"新建数据源"选项



图 1-27 使用数据源向导

数据源向导

选择如何定义连接

您有众多方式可以选择，以便数据源对其连接字符串进行定义。

基于现有连接或新连接创建数据源(C)

数据连接(D):

数据连接属性(T):

属性	值

新建(E)... 删除(L)

基于另一个对象创建数据源(A)

必须选择一个有效连接。

< 上一步(B) 下一步(N) > 完成(F) >>| 取消

图 1-28 选择如何定义连接

图 1-29　设置连接管理器

（8）在弹出的“连接测试成功”页面中，单击“确定”按钮，如图 1-30 所示。

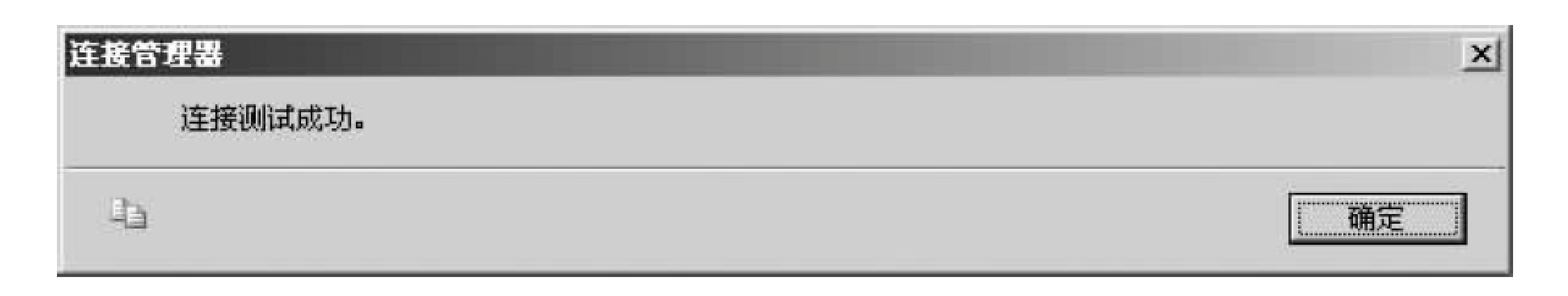


图 1-30　连接测试成功页面

（9）返回到“连接管理器”对话框，单击“确定”按钮，如图 1-31 所示。

图 1-31　返回到“连接管理器”对话框

（10）在弹出的“选择如何定义连接”页面中，单击“下一步”按钮，如图 1-32 所示。



图 1-32　选择如何定义连接

（11）在弹出的“来自现有对象的数据源”页面中，选择“基于 Analysis Services 项目创建数据源”单选按钮，单击“下一步”按钮，如图 1-33 所示。



图 1-33　选择“基于 Analysis Services 项目创建数据源”

（12）在弹出的“模拟信息”页面中，选择“使用服务账户”单选按钮，单击“下一步”按钮，如图 1-34 所示。

（13）在弹出的“完成向导”页面中，选择“使用服务账户”，单击“完成”按钮，如图 1-35 所示。

数据源向导

模拟信息
可以定义 Analysis Services 使用何种凭据来连接到数据源。

使用特定用户名和密码(S)
用户名(U):
密码(P):
使用服务帐户(V)
使用当前用户的凭据(D)
默认值(F)

< 上一步(B)　下一步(N) >　完成(F) >>|　取消

图 1-34　设置模拟信息



图 1-35　完成新建数据源

1.3.3　新建数据源视图

新建数据源视图的操作步骤如下：

(1) 建立数据源之后，建立数据源视图，在数据源视图上右击，在弹出的菜单中选择“新建数据源视图”命令，如图 1-36 所示。

(2) 在弹出的“数据源视图向导”对话框中，单击“下一步”按钮，如图 1-37 所示。

(3) 在弹出的“选择数据源”页面中，选择刚刚建立的数据源“延期纳税”项，单击“下一步”按钮，如图 1-38 所示。



图 1-36　选择新建数据源视图



图 1-37　使用数据源视图向导



图 1-38　选择数据源

（4）在弹出的“名称匹配”页面中，单击“下一步”按钮进行名称匹配，如图 1-39 所示。



图 1-39　进行名称匹配

（5）在弹出的“选择表和视图”页面中，把可选对象中所有的数据选到包含的对象中，单击“下一步”按钮，如图 1-40 所示。



图 1-40　选择表和视图

（6）在“完成向导”对话框中，单击“完成”按钮，如图 1-41 所示。

（7）得到建立的数据源视图，如图 1-42 所示。

1.3.4　浏览数据

（1）选中新建的“延期纳税批件”表，在右键菜单中选择“浏览数据”命令，如图 1-43 所示。

数据源视图向导

完成向导

请提供一个名称，然后单击“完成”以创建新数据源视图。

名称(A):

延期纳税

预览(P):

延期纳税
- dbo.税务机关代码
- dbo.延期纳税批件
- dbo.征收项目代码

< 上一步(B)　下一步(N) >　完成(F)　取消

图 1-41　完成向导



图 1-42　新建数据源视图完成

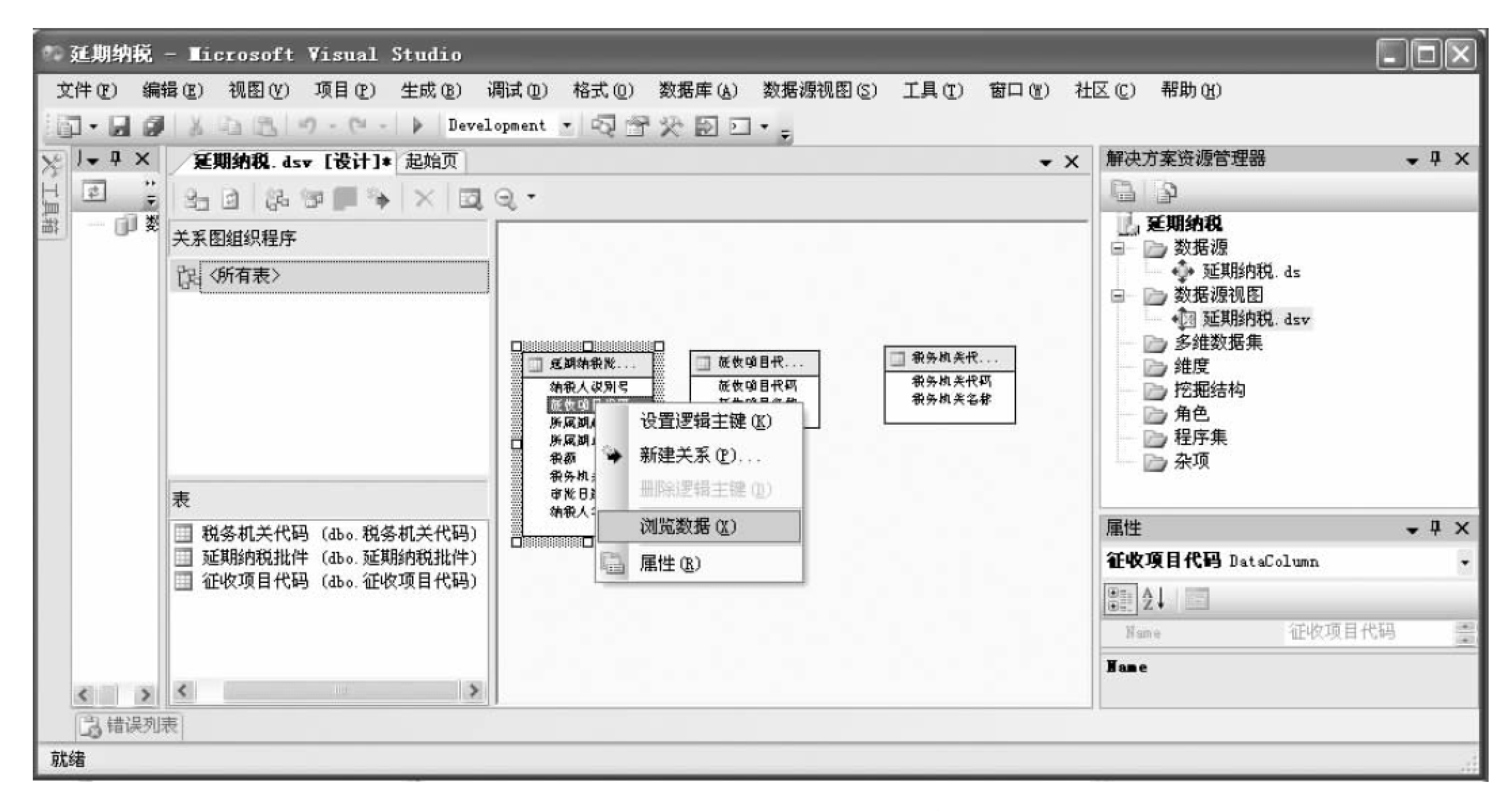


图 1-43　选择浏览数据

（2）进行数据的浏览，并且可以切换到“透视表”、“图表”、“透视图”，如图 1-44～图 1-47 所示。



图 1-44　进行数据浏览

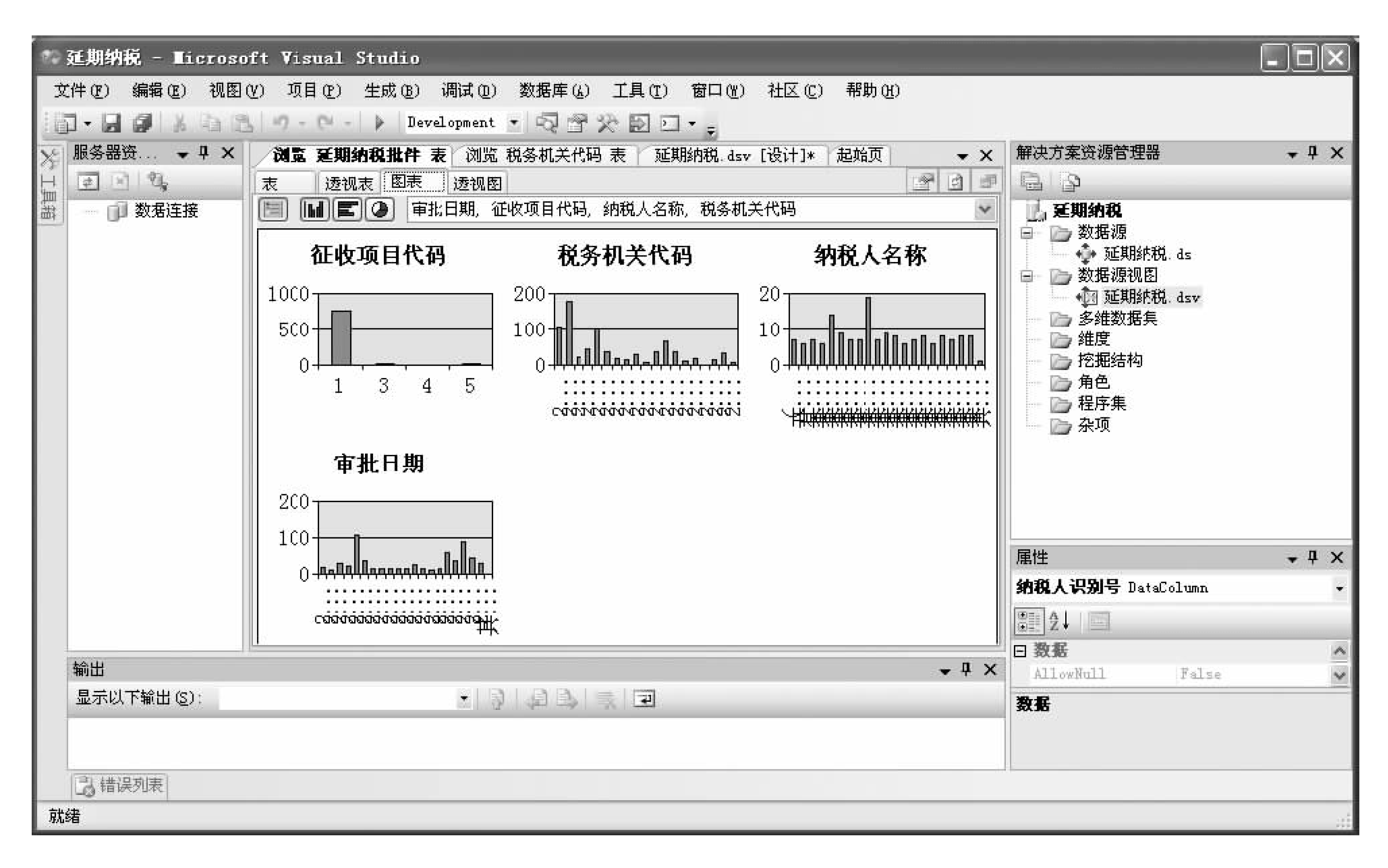


图 1-45　切换到透视表

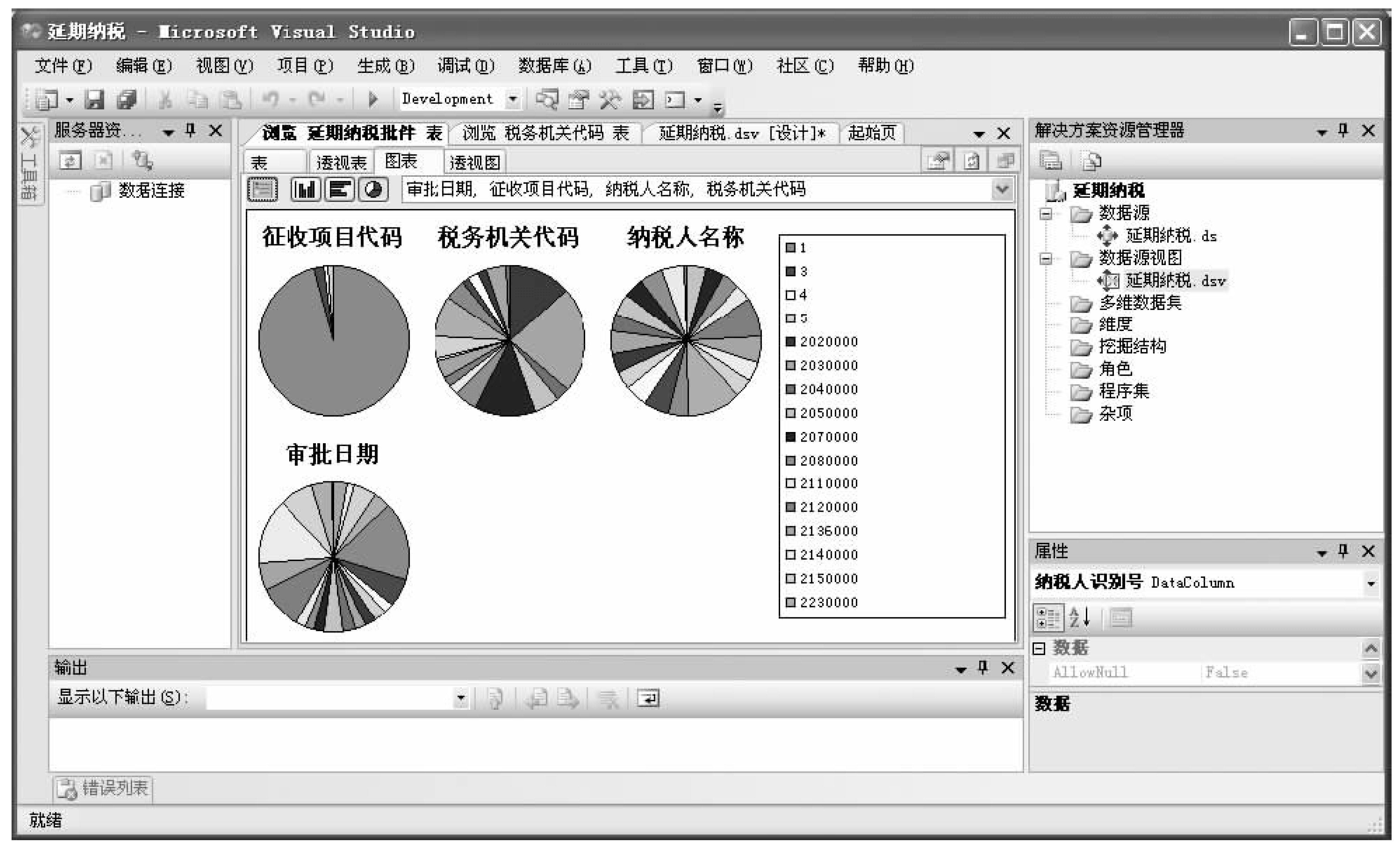


图 1-46　切换到图表

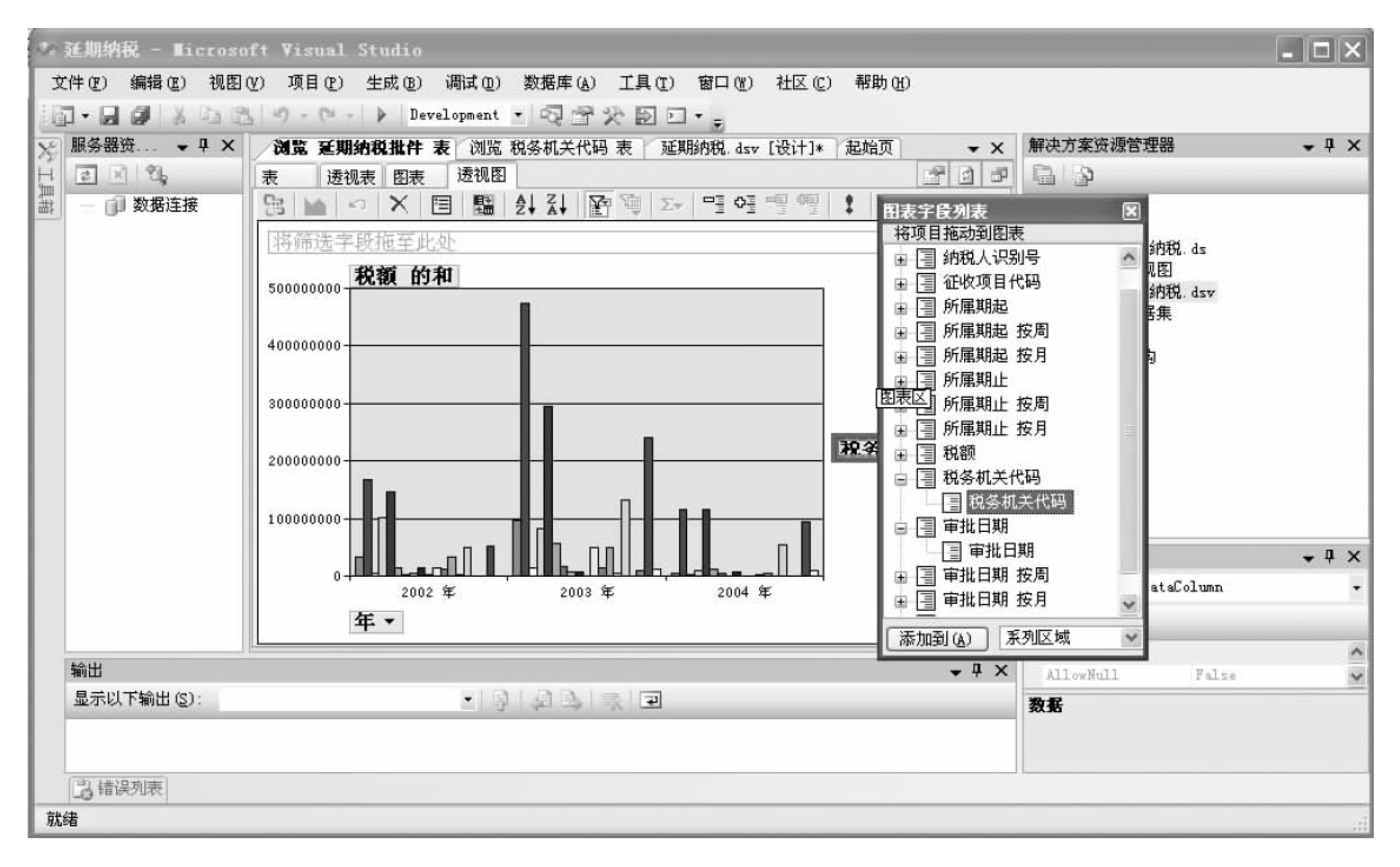


图 1-47　切换到透视图

1.3.5　数据分析

数据分析的步骤如下：

(1) 对年份数据进行趋势分析，在透视图中把“审批时间”,“税额”两个维度拖曳到表中，如图 1-48 所示。

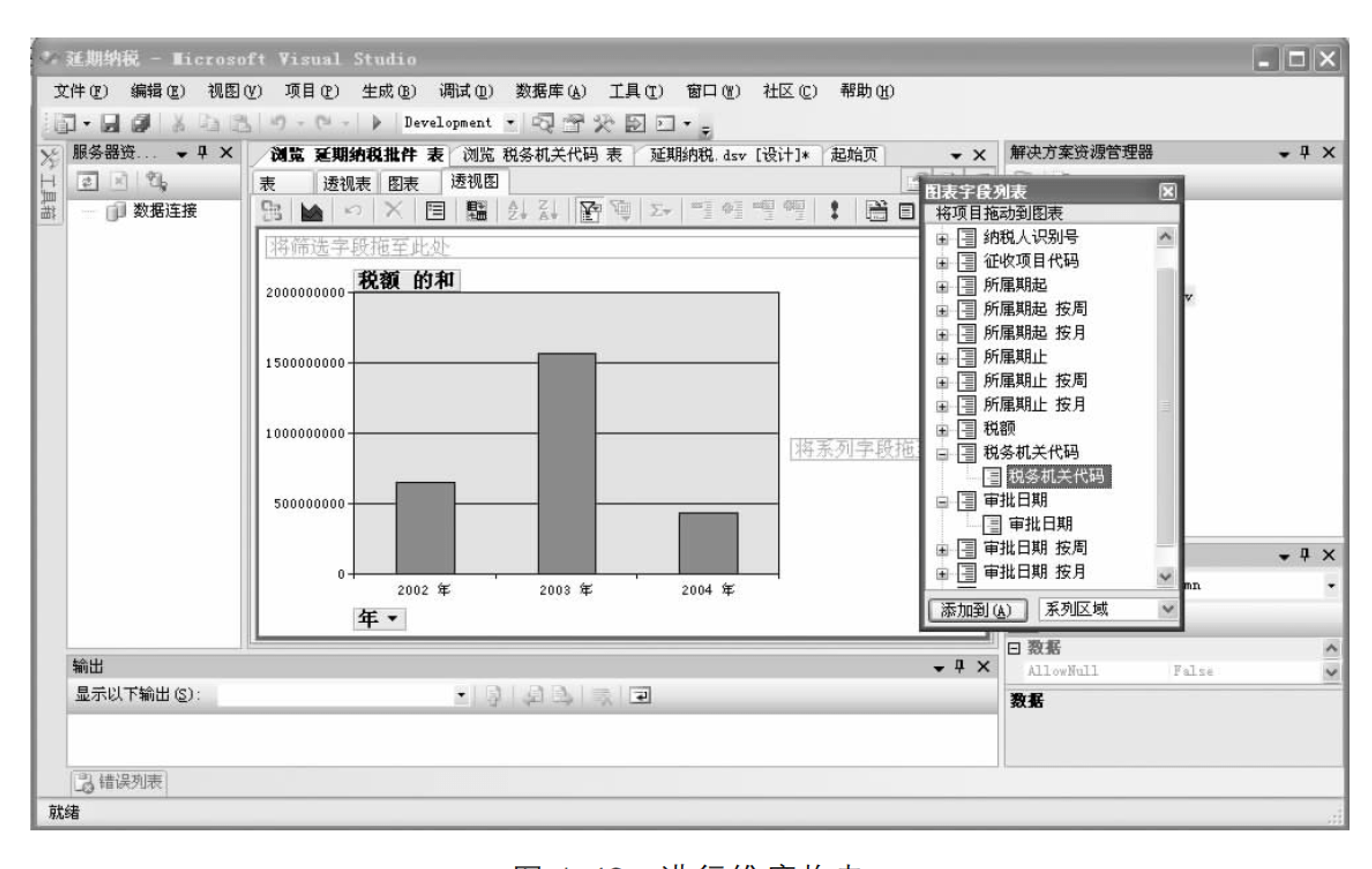


图 1-48　进行维度拖曳

由于该项目的审计时间是 2004 年 3 月，所以 2004 年数据显示的是 2004 年 1 月和 2 月的延期纳税审批金额，因此暂不将 2004 年的数据与 2002 年和 2003 年的数据进行比较。从显示的数据可以看出，2002 年全市国税系统共审核批准延期纳税 648 767 250.50 元，而

2003 年则审批 1 561 835 634.09 元，是 2002 年的两倍多。因此，应重点审计 2003 年的审批情况。

(2) 对月份数据进行趋势分析，把透视图中"审批时间"确定到 2003 年，并且按月份拖入表中，如图 1-49 所示。

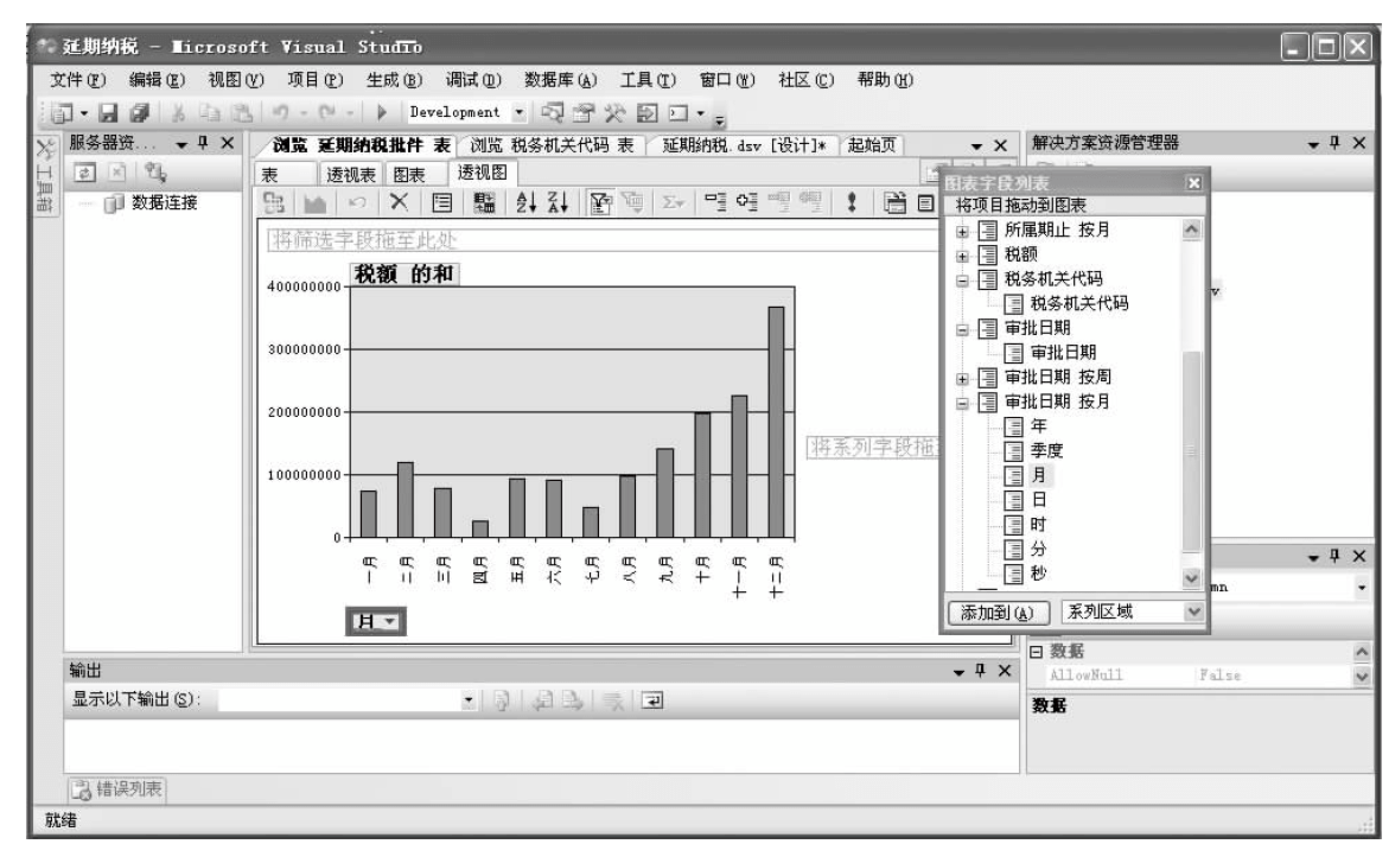


图 1-49 对审批时间数据下钻

从图 1-49 中可以看出，审批行为主要集中在年底。从 2003 年 9 月开始一直到 12 月，审批金额明显高于当年的其他月份。经过简单计算即可得出年底 4 个月的审批金额占到全年金额的 60%。对 2002 年的数据进行同样的分析也可以得出类似的结论，9～12 月的审批金额占到全年金额的 90%。

我国现行的税收工作考核机制实行的是"基数加增长比率"的传统方法。如果税务部门当年的入库税收超过了上级下达的收入任务，那么下一年的任务将在实际入库的基础上继续按照一定的比率增长，从而给税务部门形成较大的压力。因此，一些税收任务完成较好的税务部门为了使入库税收不至于过多超过收入任务，往往在年底人为调节税收收入进度，给纳税人批准延期纳税是常用的调节手段之一。

根据上述审计经验，并结合建立的多维数据集对月份数据的分析，某市国税局在被审计年度内如此集中地于年底审批延期纳税引起了审计人员的关注。

(3) 对各县区国税局的审批情况进行比较分析，把"税务机关代码"这一维度拖曳到表中，如图 1-50 所示。该市下辖区县众多，对各个区县的国税局逐一进行审计是不可能的，如何确定重点审计地区是审前调查阶段的一项重要任务。

(4) 从图中可以得到税务代码为 2030000、2070000、2270000、2810000 这 4 个地区的国税局审批的延期纳税金额远远高于其他区县，如图 1-51 所示，可对应得到 G、L、N、P 这 4 个区的国税局审批的延期纳税金额远远高于其他区县，因此可以把它们作为重点审计对象。

(5) 根据经验对数据进行其他分析，在上述基础上把"审批时间"确定为 2003 年的各个月份，"税务机关代码"确定为 2030000、2070000、2270000、2810000，把它们拖入表中，如图 1-52 所示。

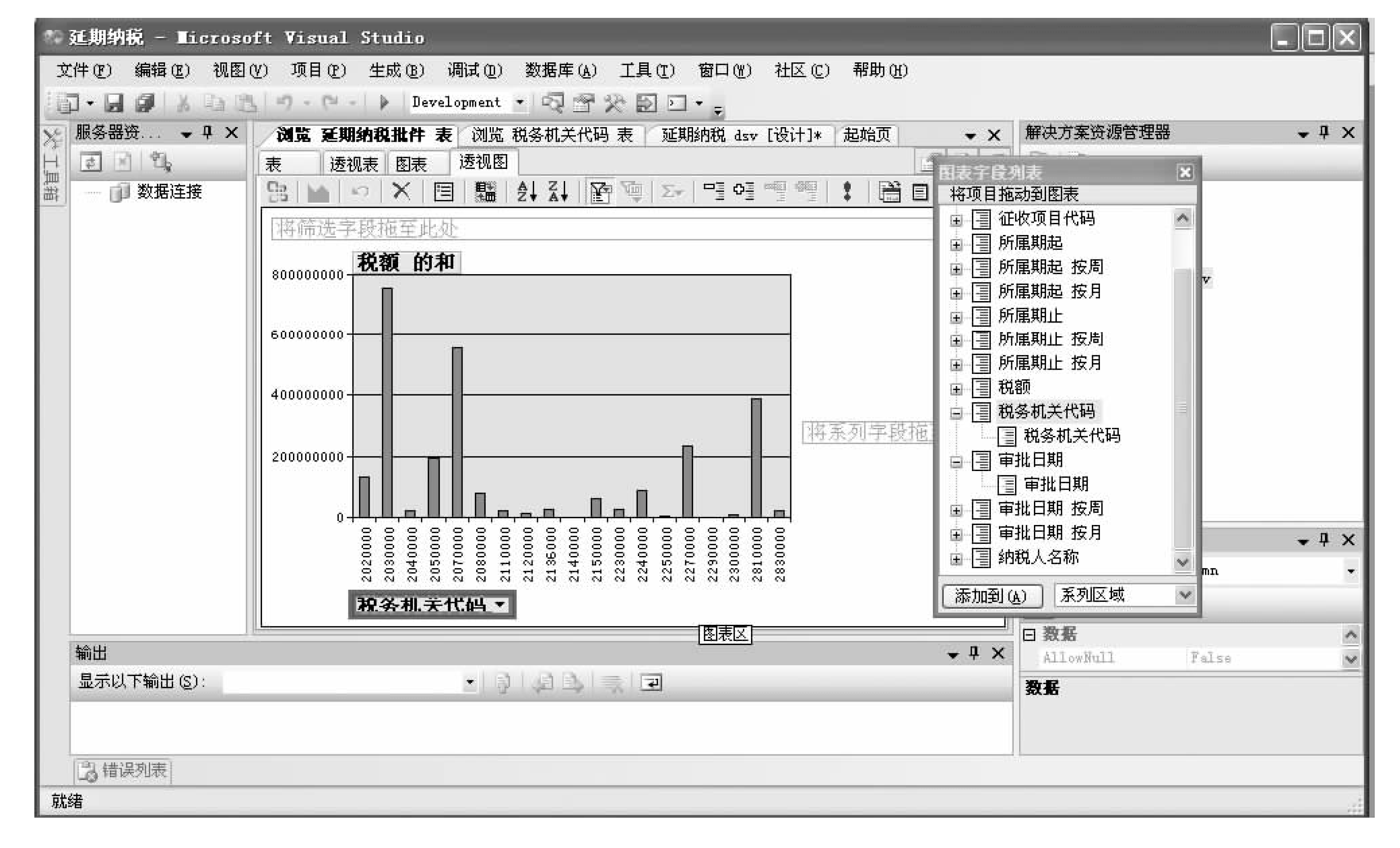


图 1-50　进行维度拖曳

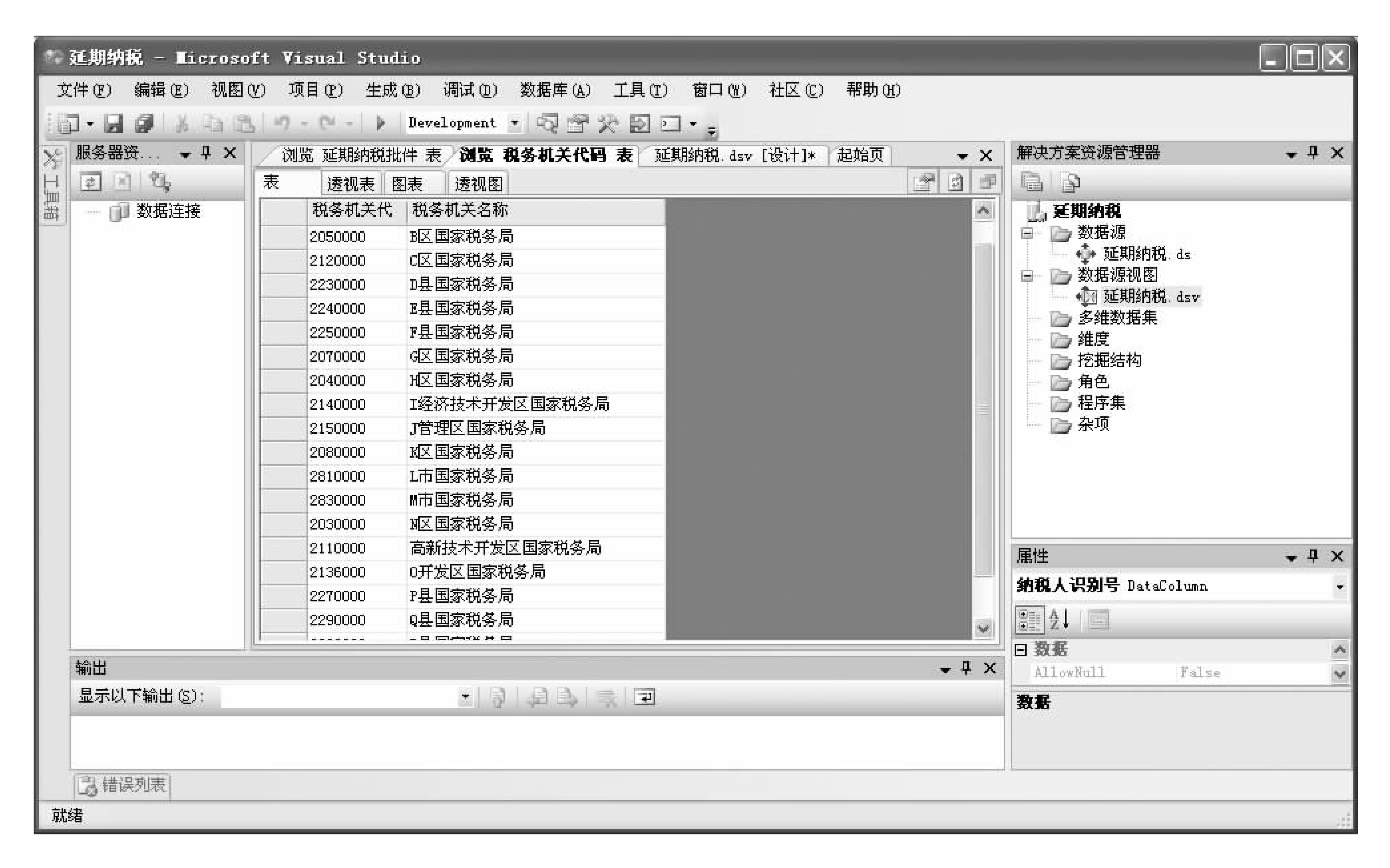


图 1-51　确定重点审计对象

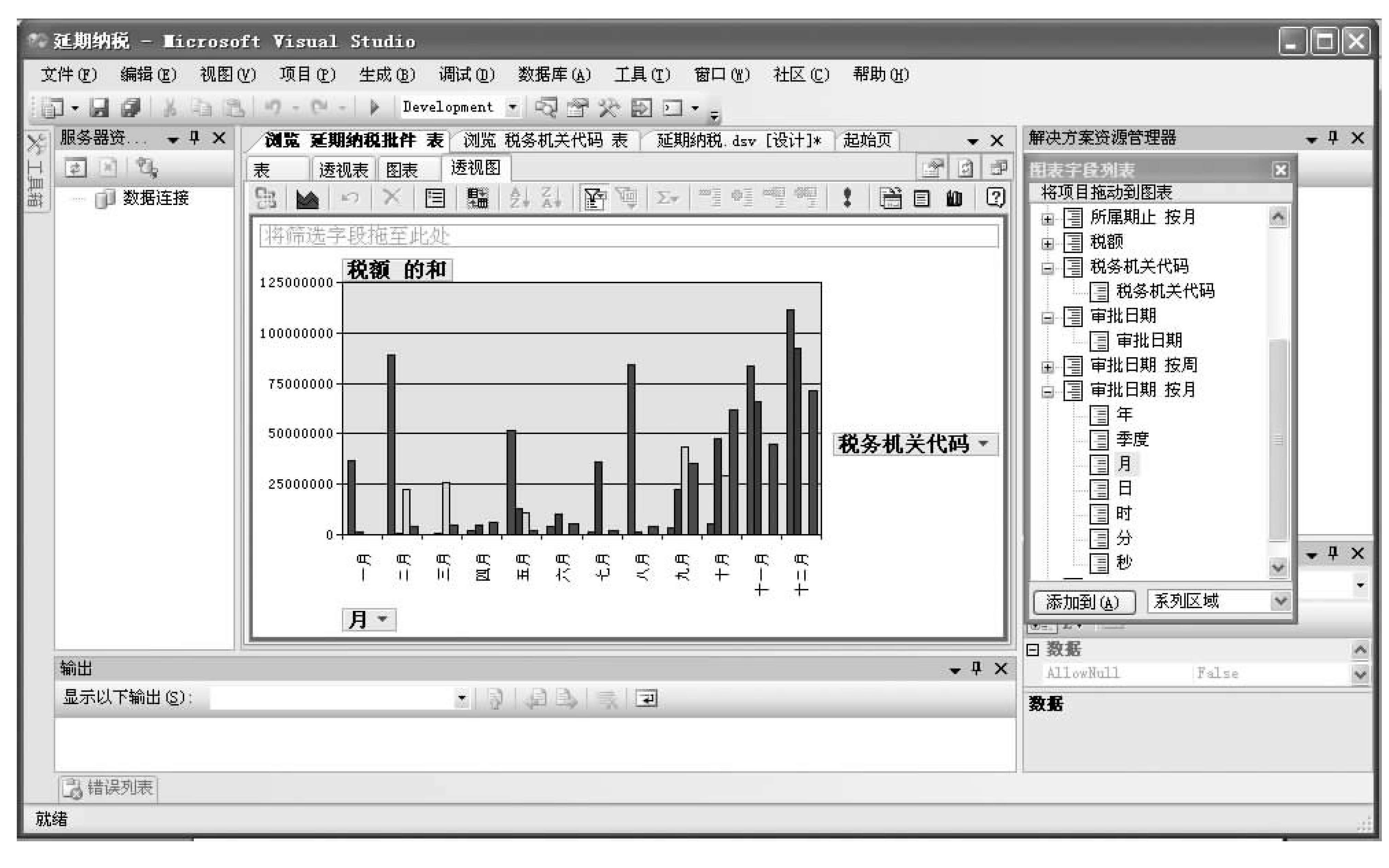


图 1-52　进行维度拖曳

（6）经过观察分析，发现税务机构代码为 2810000 在 2～7 月连续出现税务额为整数的情况，所以把“税务机关代码”选择为 2810000，继续分析，如图 1-53 所示。

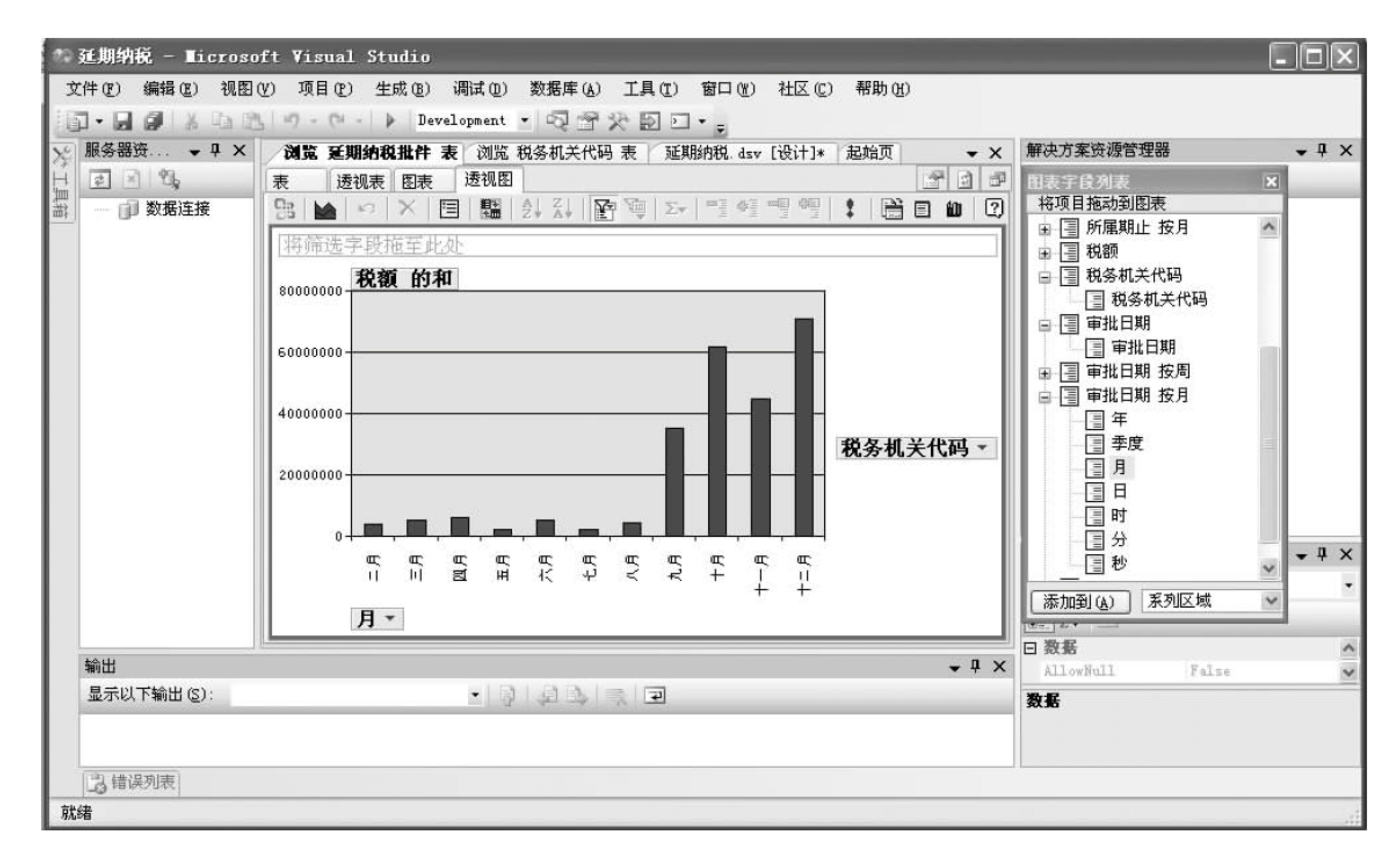


图 1-53　进一步确定审计重点

由经验可得纳税人应该缴纳的税额很少有整数的情况出现，分析至此，该市某些国税部门批准企业延期纳税、人为调节税收入库的可能性比较大。

由上述分析可得出代码为 2810000 的 L 区国家税务局从 2003 年 2～4 月连续出现审批金额为整数的情况，引起了审计人员的重视，审计重点进一步得到了明确。

1.4　案例总结

本章上述案例中，首先针对性地获得了电子数据，然后通过事实数据表的选取和维度的建立构建了多维数据集，在建立总体分析模型的基础上结合审计经验从多个角度对数据进行了分析。从统揽全局、把握总体开始，观察趋势、选择重点、运用钻取、掌握明细，最后发现线索，找到重点突破口，引导延伸，这些是本案例的基本思路和操作过程。

本案例中，始终把需求作为主线和重点，随着多维数据集的构建和分析的逐步深入，一步一步确定审计重点、缩小审计范围。在这个审计案例中，多维联机分析处理审计不仅仅体现为一种先进的技术和方法，更是作为一种思维方式在审计的整个过程中得到了贯穿和体现。

实例 2　基于关联规则方法的网上交易服务质量评价分析

2.1　任 务 描 述

科技变革与信息技术的发展,使得产品与服务在线交易成为企业与消费者无法回避的选择。到 2010 年,中国网上购物市场的交易额已达到 1800 亿元。在线交易市场具有极好的发展前景,已成为传统消费方式的一种替代。但网络虚拟空间的特性,无疑提高了消费者交易的风险和感知服务质量评价的难度,也加大了企业服务质量改进的难度。

在互联网环境下,哪些因素会决定着服务质量水平呢? 某研究机构罗列了 29 个有可能影响服务质量水平的因素和 1 个决策属性“顾客情绪不佳”。这 30 个因素为:

- 性别(xb);
- 商家称网络系统有误(shjxtyw);
- 顾客感觉网络系统有误(gkxtyw);
- 商家突然涨价(shjzj);
- 网络价格与实收价格不一致(jgbyz);
- 服务过程中乱收费(fwlshf);
- 产品质量问题(zhlwt);
- 服务不讲诚信(fwbchx);
- 误导消费者(wdxfzh);
- 商家称工作失误(shjgzshw);
- 顾客感觉工作失误(gkgzshw);
- 服务态度差(fwtdch);
- 不合理的规定(霸王条款)(bhlgd);
- 不能在网站查询交易的步骤(bnchxjybzh);
- 企业对交易的反应速度慢(fysdm);
- 企业对交易的反应天数(fytsh);
- 退货后不能成功退款(bchgtk);
- 退款的速度慢(tksdm);
- 退款天数(tktsh);
- 对投诉的处理结果糟糕(tsjgz);
- 对投诉的处理速度慢(tschlm);
- 投诉处理天数(tschltsh);
- 客服打不通(kfdbt);
- 客服态度差(kftdch);
- 发票速度慢(fpsdm);

- 开发票天数(kfptsh)；
- 发票(行程单)错误(fpcw)；
- 多次修改仍不对(dcxgbd)；
- 产品等被擅自更换(chpbgh)；
- 顾客情绪不佳(gkqqbj)。

另外，此研究机构还搜集了 1366 个顾客的数据，部分原始数据如表 2-1 所示，请根据这些数据(服务质量数据 1. xls)进行以下分析。

表 2-1 部分原始数据

编号	性别（男1女2）	商家称网络系统有误	顾客感觉网络系统有误	商家突然涨价	网络价格与实收价格不一致	服务过程中乱收费	产品质量问题
TS773717	2	0	0	0	0	0	0
TS773385	2	0	0	0	0	0	0
TS773321	2	0	0	0	0	0	0
TS772923	1	0	0	0	0	0	1
TS772854	1	0	0	0	0	0	0
TS772833	1	0	0	0	0	0	0
TS772674	2	0	1.5	0	0	0	1.5
TS772216	2	0	0	0	0	1	0
TS772202	2	0	1	0	0	0	0
TS772193	1	0	0	0	1.5	0	0
TS772066	2	0	0	0	0	0	0
TS771997	2	1	0	0	0	0	0
TS771266	1	0	0	0	0	0	0
TS771118	1	0	0	0	0	0	0
TS771050	1	0	0	0	0	0	0
TS771030	2	0	0	0	0	0	0
TS770768	1	0	0	0	0	0	1
TS770170	1	0	0	0	0	0	0
TS769914	1	0	0	0	0	0	2.5

(1) 找出 29 个因素中影响服务质量的主要因素。

(2) 找出主要影响因素和服务质量水平之间的关联规则。

2.2 技术原理

2.2.1 关联规则的概念

关联规则是形如 $A \Rightarrow B$ 的蕴涵式。规则 $A \Rightarrow B$ 的支持度 $s(A \Rightarrow B)$ 定义为 D 中包含 $A \cup B$ 的事务所占的百分比，表示项集 $A \cup B$ 在 D 中出现的概率。规则 $A \Rightarrow B$ 的置信度 c 定义为 D 中包含项集 $A \cup B$ 的事务数和包含项集 A 的事务数的比值，表示当项集 A 出现时，项集 B 出现的概率，置信度大于用户指定的最小置信度值的规则是可信的。

关联规则 computer => antivirus_software (support = 2%, confidence = 60%)，表示 2%的顾客同时购买计算机和杀毒软件，购买计算机的顾客 60%也购买了杀毒软件。

2.2.2 Apriori 算法

Apriori 的命名是因为算法使用了频繁项集性质的先验知识，即 Apriori 性质。Apriori 性质的内容是：频繁项集的所有非空子集也都必须是频繁的。此性质被用于减少候选频繁项集的数量。Apriori 算法将发现关联规则的过程分为两步：第 1 步是通过迭代，检索出源数据中的所有频繁项集，即支持度不低于用户设定阈值的项集；第 2 步是利用第 1 步中检索出的频繁项集构造出满足用户最小信任度的规则。

对于如表 2-2 所示的数据集，产生频繁项集的过程如图 2-1 所示。

表 2-2 某销售数据集

TID	商品 ID 的列表	TID	商品 ID 的列表	TID	商品 ID 的列表
T100	I1,I2,I5	T400	I1,I2,I4	T700	I1,I3
T200	I2,I4	T500	I1,I3	T800	I1,I2,I3,I5
T300	I2,I3	T600	I2,I3	T900	I1,I2,I3

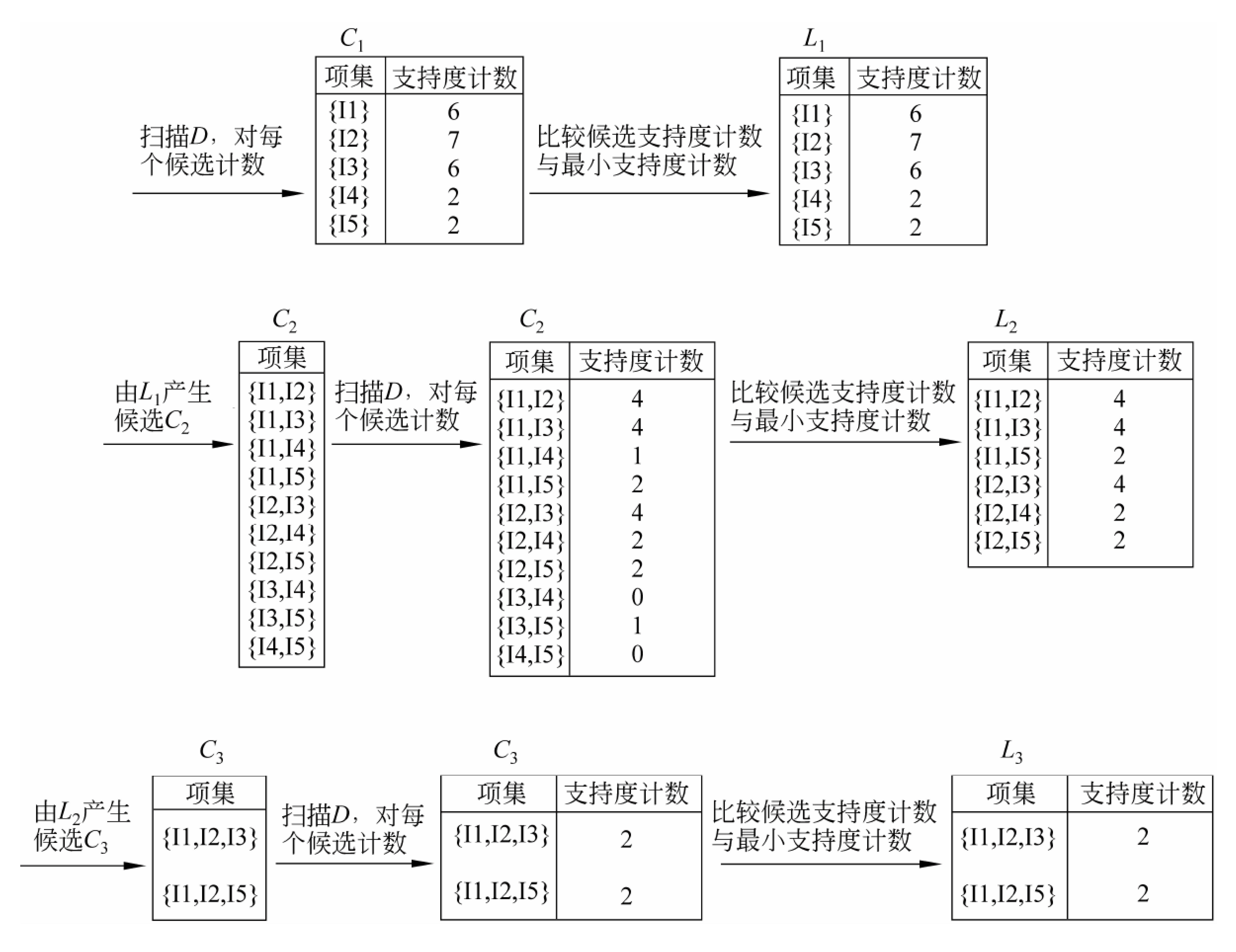


图 2-1 产生频繁项集示例

2.3 具体实现

（1）通过对影响因素的初步解读，有些因素表示的信息重复，如企业对交易的反应速度慢（fysdm）和企业对交易的反应天数（fytsh）、退款的速度慢（tksdm）和退款天数（tktsh）、对投诉的处理速度慢（tschlm）和投诉处理天数（tschltsh）、发票速度慢（fpsdm）和开发票天数（kfptsh）。于是可将下列表示重复信息的因素删除：

- 企业对交易的反应天数（fytsh）；
- 退款天数（tktsh）；
- 投诉处理天数（tschltsh）；
- 开发票天数（kfptsh）。

（2）将影响因素的中文表示替换成英文表示；将缺失数据用 null 进行填充；将属性值进行以下离散化，预处理后的数据存储在文件“服务质量数据 2.csv”中。

- 若属性值为 0，则将属性值替换为 a；

- 若属性值大于 0 且小于等于 1，则将属性值替换为 b；
- 若属性值大于 1 且小于等于 2，则将属性值替换为 c；
- 若属性值大于 2 且小于等于 3，则将属性值替换为 d。

（3）选择“开始”→“所有程序”→Weka3.6.5→Weka3.6 命令，如图 2-2 所示。

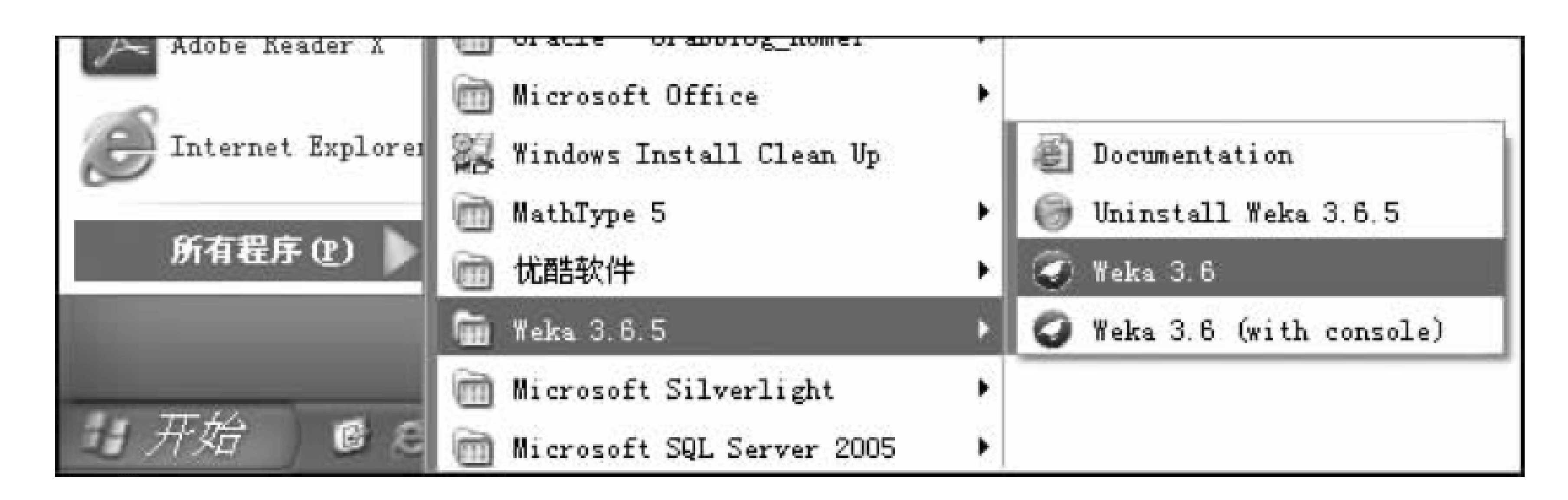


图 2-2　打开 Weka 软件

（4）单击 Explorer 按钮，如图 2-3 所示。

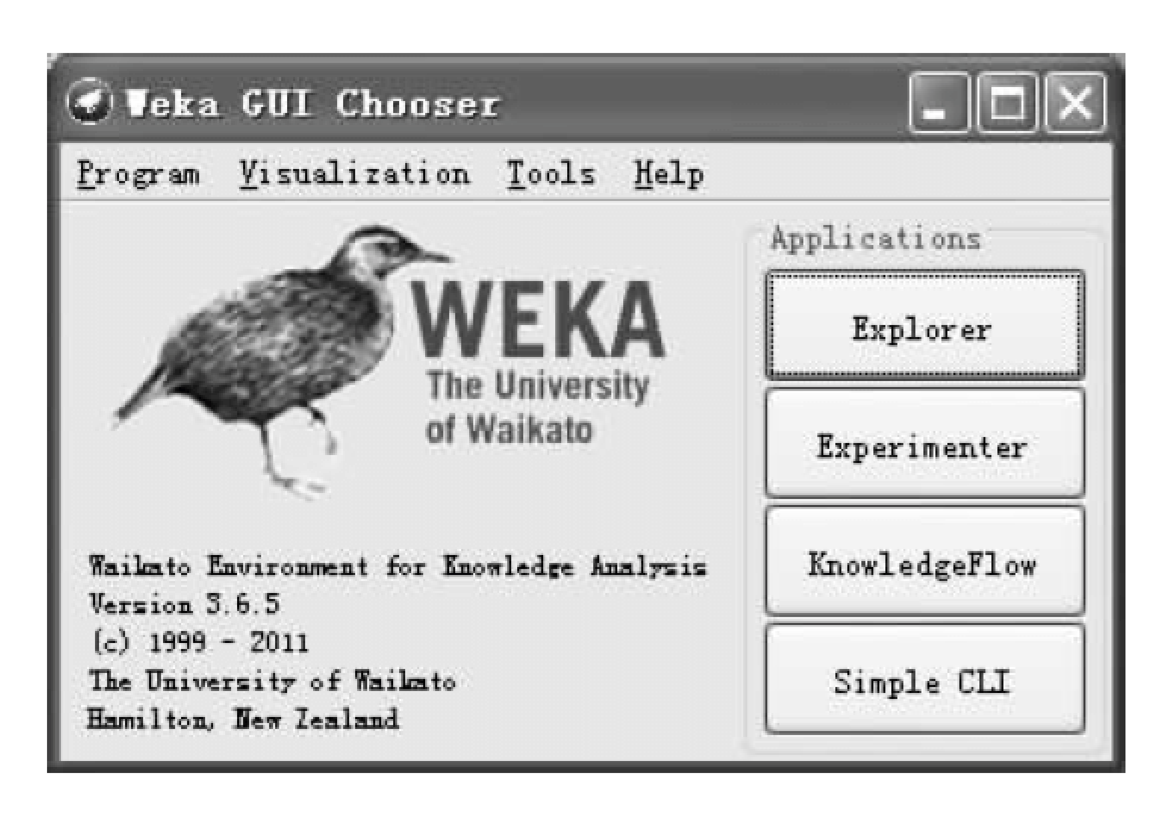


图 2-3　打开 Explorer 应用

（5）单击 Open file 按钮，选择要打开的文件“服务质量数据 2.csv”。单击“打开”按钮，如图 2-4 所示。



图 2-4　打开数据文件

（6）在如图 2-5 所示的界面中，可以知道“服务质量数据 2”数据集中共有 1366 个实例，每个实例有 26 个属性。选中某个属性，可以查看 1366 个实例关于这个属性的属性值取值信息。

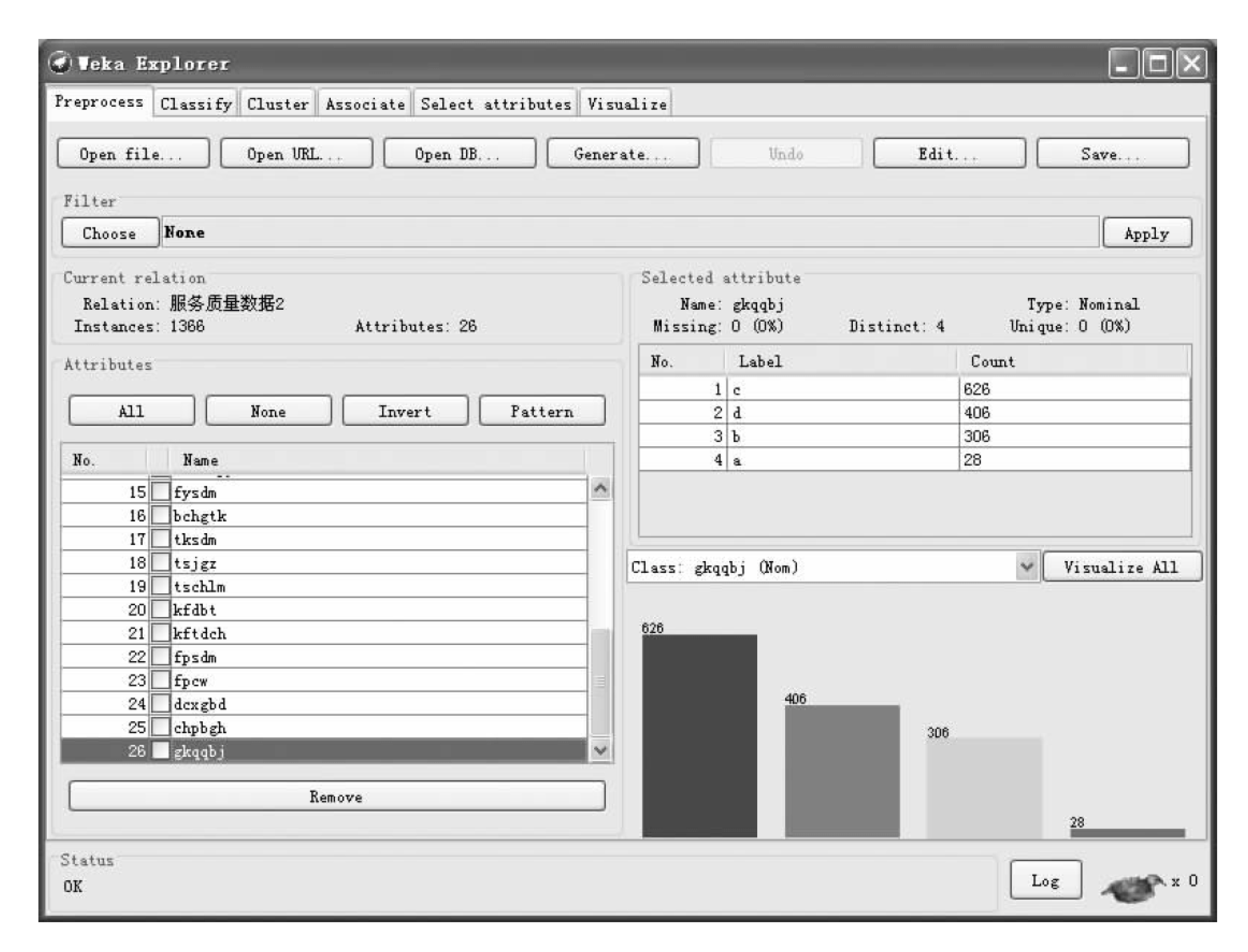


图 2-5　查看数据特征

（7）单击 Select Attributes 标签，在 Attribute Evaluator 栏中选择 SfsSubsetEval 项，单击 Close 按钮，如图 2-6 所示。

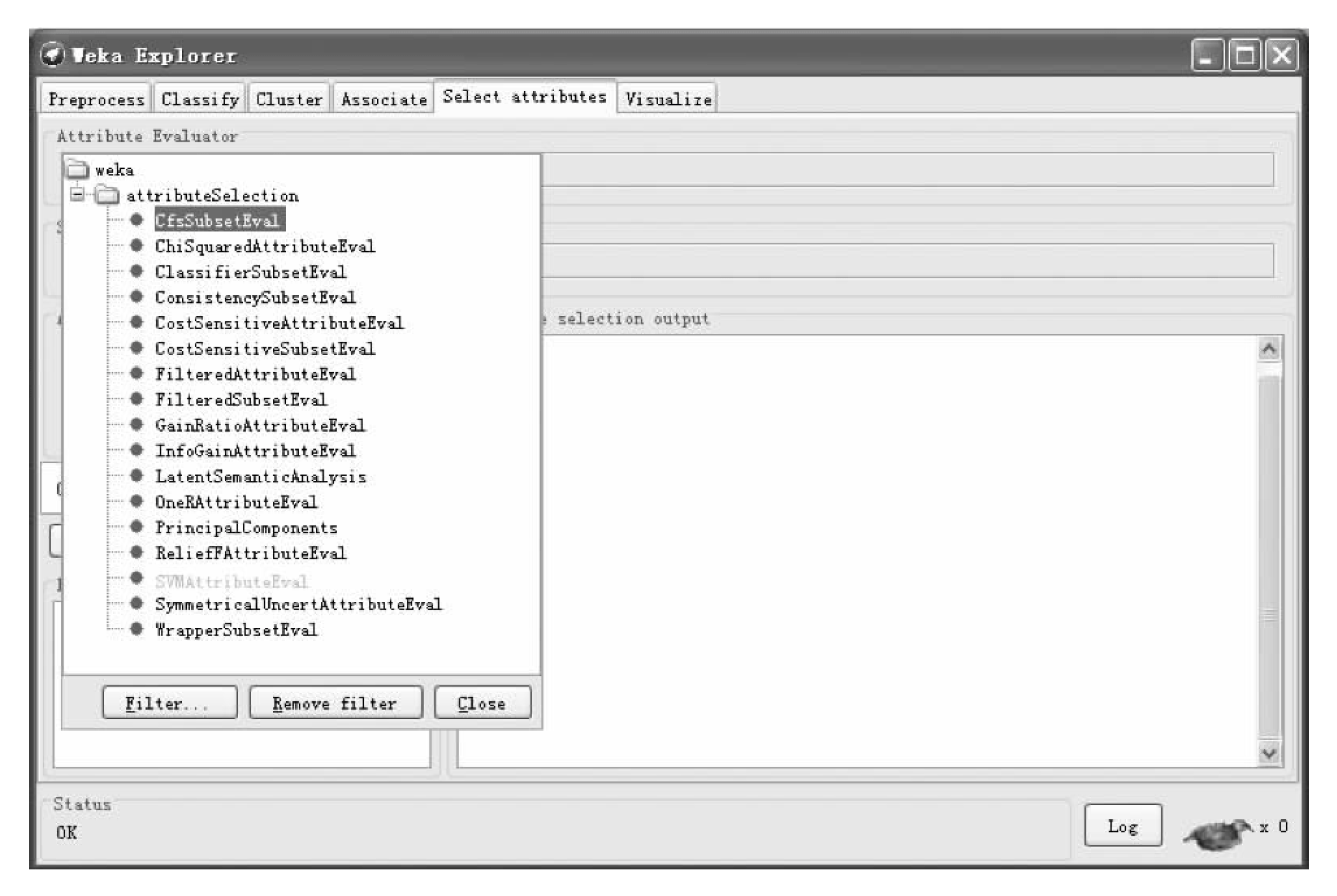


图 2-6　选择属性选取方法

（8）在 Search Method 栏中选择 BestFirst 项并单击 Close 按钮，如图 2-7 所示。

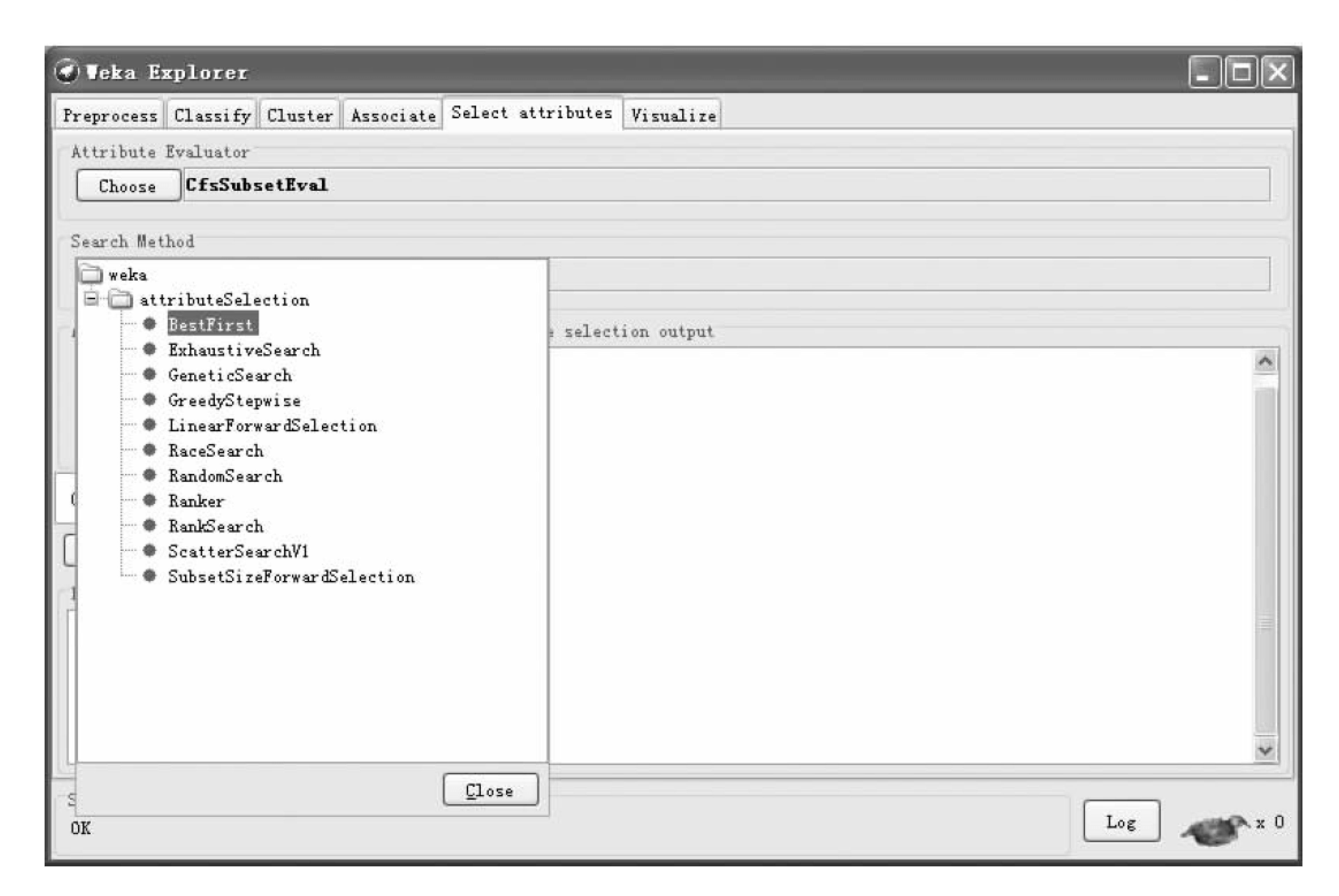


图 2-7　选择搜索方法

（9）单击 Start 按钮，运行结果如图 2-8 所示，可知在 25 个因素中，fwbchx、gkgzshw、fwtdch、bnchxjybzh、fysdm、kftdch 六个因素为重要因素。

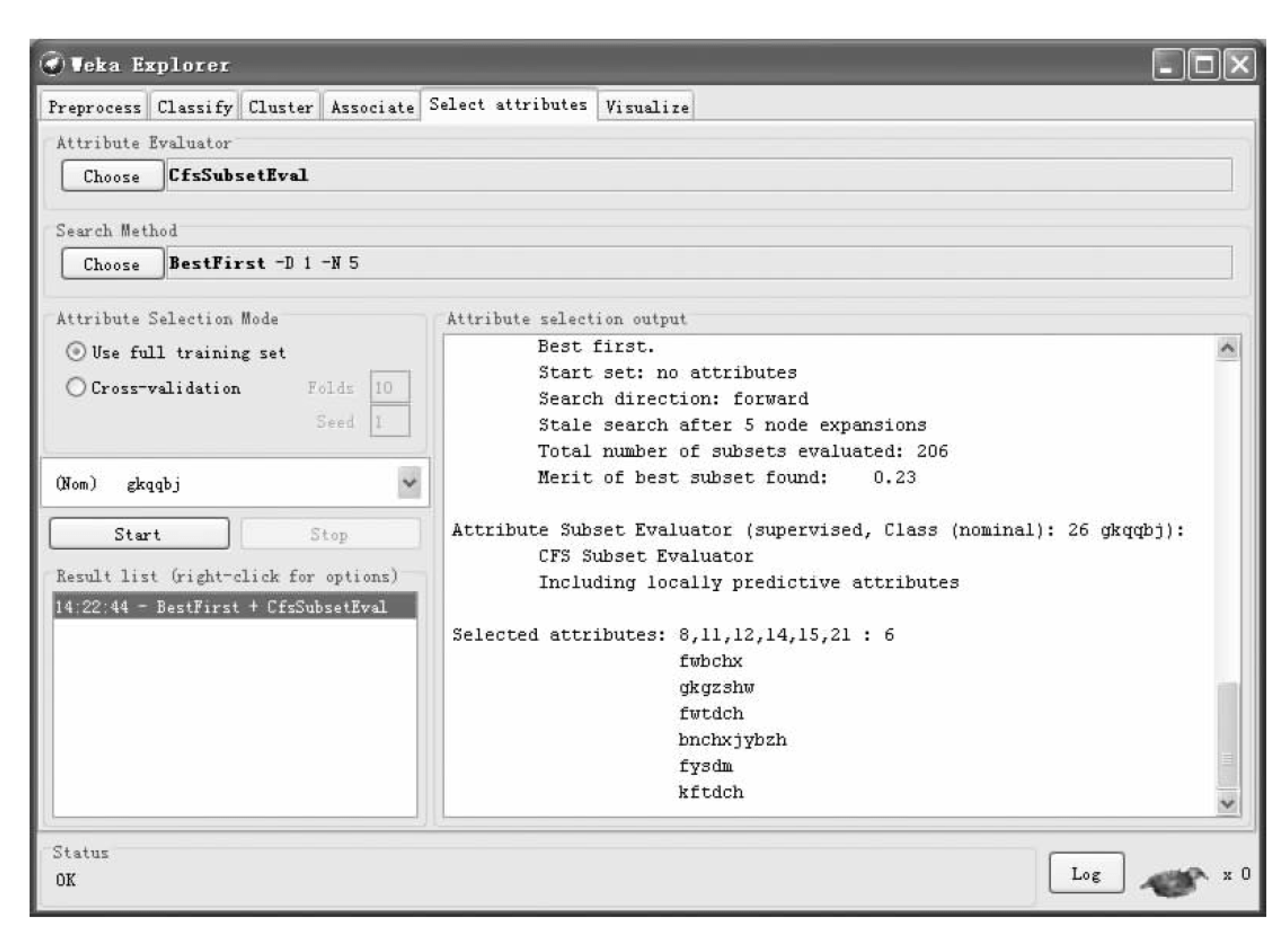


图 2-8　得到重要属性

（10）对“服务质量数据 2”数据表进行处理，只保留以上 6 个影响因素和 1 个决策属性，得到“服务质量数据 3”数据表。

（11）打开文件“服务质量数据 3. csv”，如图 2-9 所示。



图 2-9　根据重要属性得到新数据表

（12）单击 Associate 标签，并选择 Apriori 算法，如图 2-10 所示。

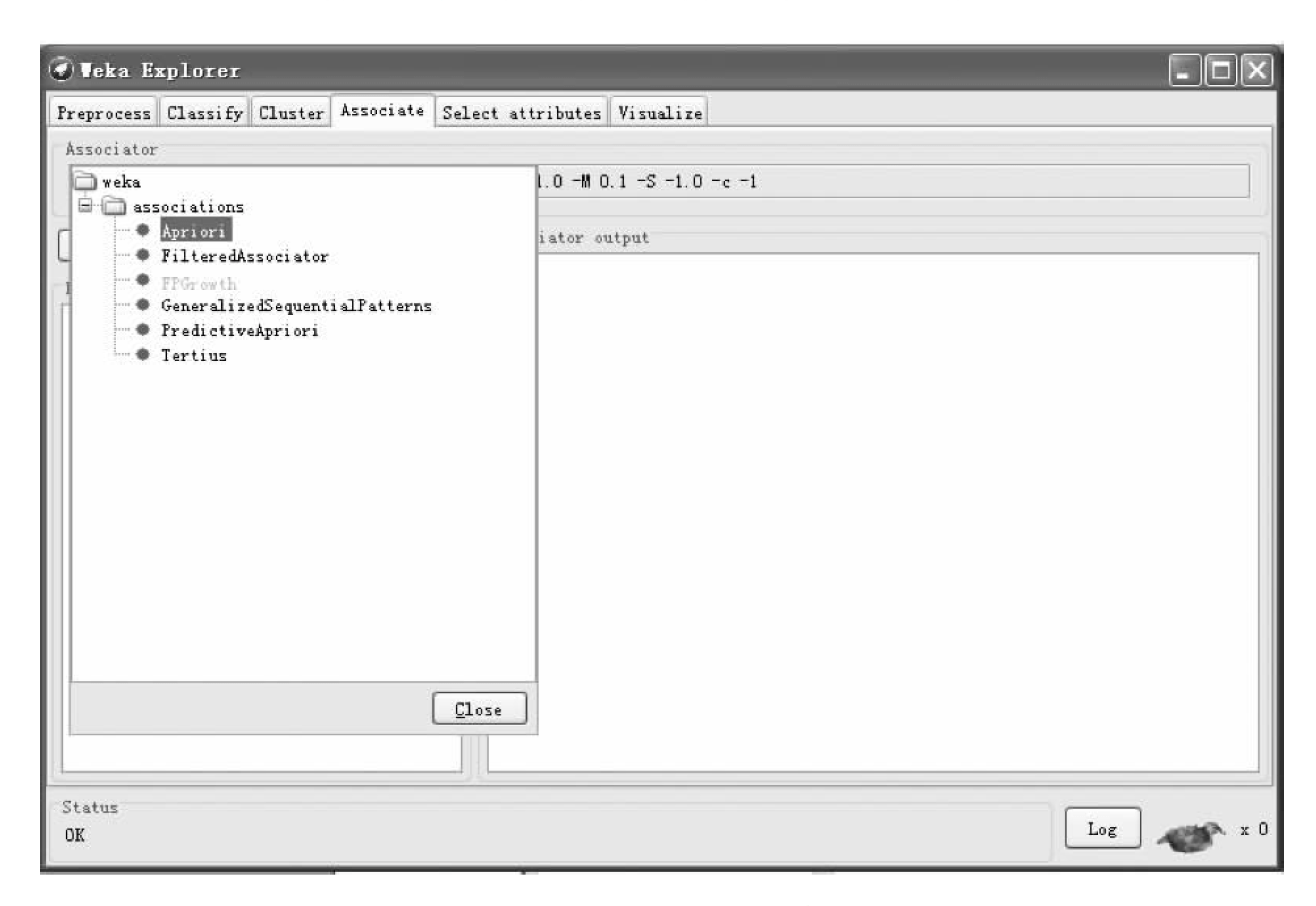


图 2-10　选择 Aprior 算法

（13）双击 Apriori 可以对算法的参数进行设置，参数设置如图 2-11 所示。

（14）单击 Start 按钮，Weka 软件显示运行结果，如图 2-12 所示。

（15）在结果显示中还可以看到，一共得到 10 条关联规则，在每条规则后附有规则的置信度。第一条规则为 fwbchx=d kftdch=d 160==>gkqqbj=d 145 conf:(0.91)。这条

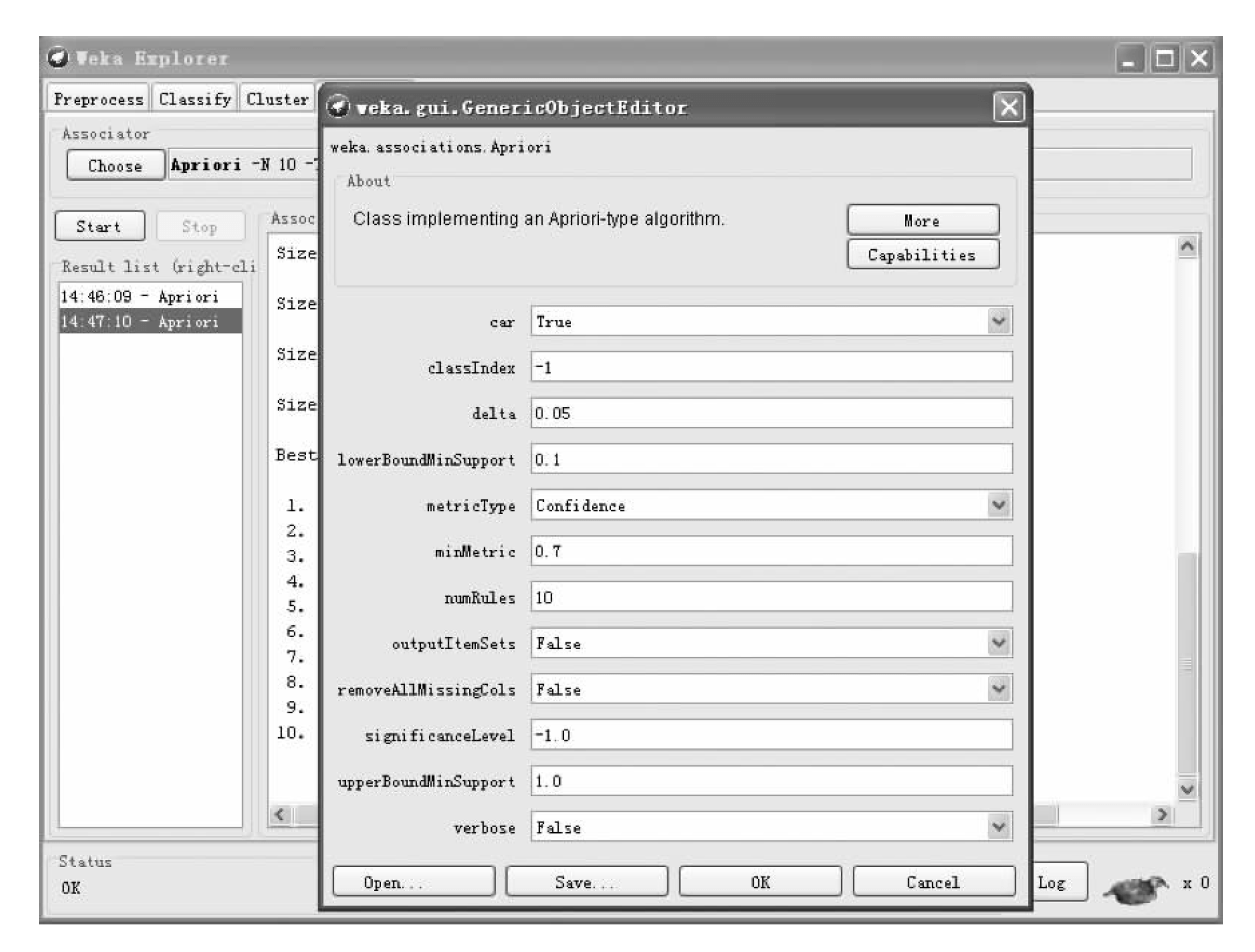


图 2-11　进行算法参数设置

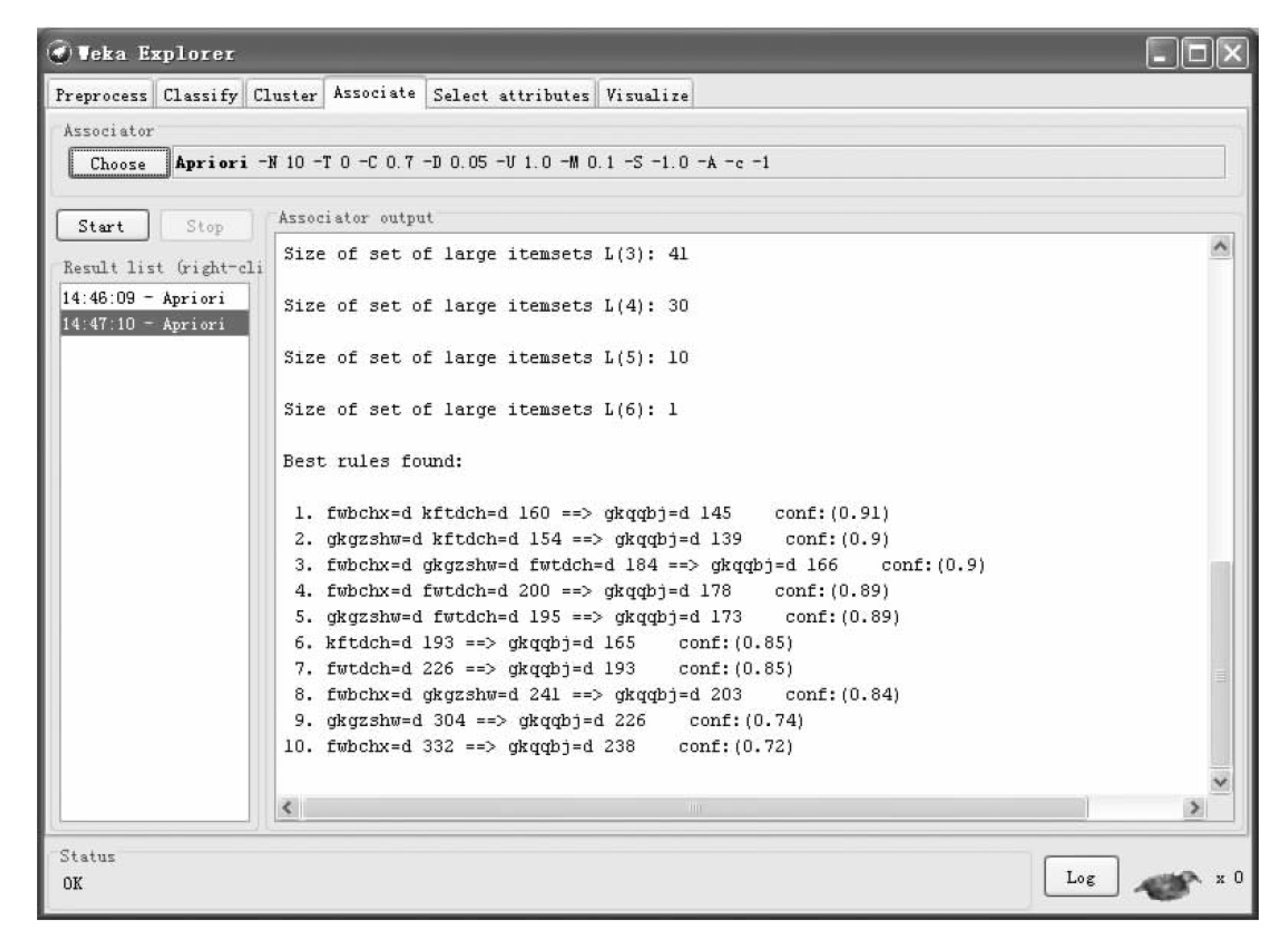


图 2-12　得到关联规则

规则可以解释为，在 1366 名客户中，有 160 名客户认为服务不讲诚信且客服态度差，其中的 145 名客户因此导致情绪不佳，即对服务质量不满意。

至此，得到了影响服务质量的主要因素以及可能导致顾客情绪不佳的关联知识，完成了

任务要求。

2.4 案例小结

在线网上交易市场具有极好的发展前景，但由于网络的一些特性，使得网上交易的服务质量评价不同于传统的服务质量评价，于是网上交易服务质量如何进行评价成为管理科学的一个研究课题。某研究机构获取了网上交易服务质量评价的数据，但是如何从这些大量的数据中抽取出关于网上交易服务质量评价的知识是关键所在。本案例使用数据挖掘中的属性选取技术和关联分析技术，利用 Weka 软件平台，成功得到了影响服务质量的主要因素以及可能导致顾客情绪不佳的关联知识，为网上交易服务质量评价研究提供了定量结果。

实例 3　基于 Weka KnowledgFlow 模块的大学生专业方向预测分析

3.1　任务描述

随着专业划分越来越细,大学生在专业发展方向上有了更多的选择余地。例如,计算机专业的大学生,可以根据自己的兴趣爱好选择计算机软件、硬件、网络、多媒体等方向。每个人的思维方式不同,对所学知识的偏爱程度和理解程度不同,在不同发展方向做出的成绩也就相去甚远。如果能够通过科学的分析手段让每个人了解自己的特点,找到最佳发展方向,或有意识地培养某方面的能力,会有助于个人发展。但是,并不是每个人都对自己的兴趣爱好或专长有充分正确的了解,有时甚至存在错误的认识,通过客观数据帮助学生发现自己的特长,就是本例所要解决的问题。

假设学生在某些科目中取得的成绩和他在相关专业方向中的能力有着密切联系,是否能够借助学生在各门基础课中所取得的成绩,预测其在哪个发展方向上将会有较好的表现,即找出基础课成绩对专业课成绩和实践成绩的影响规律呢?假设有这样的可能,就可以在学生选修专业课时,指导他们根据自己基础课的成绩情况决定选修哪方面的专业课和实践课。本实例的任务就是发现这样的规律。

下面介绍采用关联规则挖掘模型实现上述目的的过程,包括挖掘模型选择、模型训练、评估和应用。

3.2　技术原理

3.2.1　数据收集和准备

在本实例中,采用了某大学计算机科学与技术专业历届毕业生的成绩信息。由于各届学生的培养方案不同,所修科目有所不同,并且课程中包含了百分制成绩(考试)和五级制成绩(考查)两种不同的成绩形式,个别学生由于没有完成学业等原因,成绩记录不完整。这些因素使得原始数据不能直接用于数据挖掘,必须通过预处理方法将数据处理成为“干净”、统一的挖掘数据源。

3.2.2　模型选择

该问题的解决采用 Weka 的 KnowledgFlow 模块,选择其中的 Apriori 算法对数据进行分析。

3.3　具体实现

3.3.1　数据预处理

数据的原始形式如图 3-1 所示,每个班级的成绩数据保存为一个数据库表。

女	1982112932000505	计0005	5年	200020002	毛泽东思想概论	76	大学数学(一)B	78	大学语文	良
女	1979111032000506	计0005	5年	200020002	毛泽东思想概论	80	大学数学(一)B	76	大学语文	优
女	1981042532000507	计0005	5年	200020002	毛泽东思想概论	70	大学数学(一)B	92	大学语文	优
女	1977100432000508	计0005	5年	200020002	毛泽东思想概论	61	大学数学(一)B	81	大学语文	优
男	1982011132000509	计0005	5年	200020002	毛泽东思想概论	54	大学数学(一)B	60	大学语文	中
男	1980090132000510	计0005	5年	200020002	毛泽东思想概论	66	大学数学(一)B	70	大学语文	中
女	1982050332000511	计0005	5年	200020002	毛泽东思想概论	72	大学数学(一)B	84	大学语文	良
女	1981122632000512	计0005	5年	200020002	毛泽东思想概论	77	大学数学(一)B	90	大学语文	良

图 3-1　原始数据形式

显然,原始数据不能够直接产生挖掘数据源,需要采用多种方法进行预处理。下面按预处理的执行过程,依次介绍所用的预处理方法和预处理之后的效果。

1. 处理空缺值

一般地,在大学生成绩数据库中,产生空缺值的原因是由于学生中途终止学业,成绩管理系统仍然保存该生已有的成绩数据,所以产生了一些只有部分课程成绩的记录。根据挖掘目的,要用基础课成绩预测专业课成绩,这样不完整的数据对挖掘任务是没有帮助的,所以对这类空缺数据采用删除记录的方法,从挖掘数据源中去掉对应的整条记录。只保留正常毕业或者结业的学生成绩数据。

2. 属性选择

成绩数据库包含学生的部分基本信息,如姓名、班级等,这些信息与挖掘目标没有直接关系,需要去除以减少数据的维度。另外,某些课程由于其授课和考核方式的原因,可能会出现所有学生或绝大多数学生的该门课程成绩均为"良"(也许是"优"、"中"或"及格"),这样的属性不仅对挖掘没有帮助,而且可能产生虚假的规则,所以也需要从属性中去除。

3. 数据的规范化

如果挖掘算法使用离散的数据类型,则要对连续型成绩(百分制成绩)进行离散化处理。由于各种各样的原因,某些课程的成绩会表现出偏高、偏低、分散或集中等特点。例如,同样年级的 A、B、C、D 四个班在同一个学期开始课程 X,A、B 班由教师 1 任课,C、D 班由教师 2 任课,四个班的成绩分布就可能不同,这种情况下,衡量一个同学的成绩好坏就应该根据每个学生所在的班级总体成绩分布的前提下进行。如果盲目地将成绩数据进行离散化,可能会破坏原始数据中所包含的信息。

采用数据规范化的方法,将所有成绩数据规范到某一个设定的范围中(如 60～100),对所有成绩的离散化都在这个统一的成绩范围内进行,可以得到更客观的离散化结果。

设某班级某课程的原始成绩分布在 MinI 和 MaxI 之间,采用线性变换式(3-1)映射到 60～100 之间:

$$V' = \frac{V - \mathrm{MinI}}{\mathrm{MaxI} - \mathrm{MinI}}(100 - 60) + 60 \tag{3-1}$$

其中,V 是原始数据值;V' 是规范化之后的数据值。

规范化结果是使得所有数据表(对应不同班级的成绩数据)中的连续型成绩数据都映射到了一个统一的范围中。

4. 属性概化

由于每位学生在大学期间要学习几十门课程,如果每门课程作为一个属性,算法中处理

的属性较多,同时得出的规则也比较复杂,为了简化计算过程和结果表示,对属性进行概化。例如,将学生各个学期的数学课合并为一个属性,成绩取平均;物理、英语、政治等课程也做同样处理。还可以进一步对硬件基础课,如“模拟电路”、“数字电路”进行合并;对软件类基础课,如“算法与程序设计”、“数据结构”等进行合并。专业课也可以按照专业方向合并,如“计算机体系结构”、“组成原理”等合并为硬件。合并之后,大大减少了需要处理的属性,挖掘结果也能够得到简化。

5. 数据离散化

原始的 Apriori 算法只能处理布尔型数据,Weka 中的 Apriori 模型则可以处理离散型数据。本例中,成绩数据有两种形式,考试课成绩为百分制,考查课成绩按“优”、“良”、“中”、“及格”和“不及格”5 级记录,需要将数据进行离散化处理。

根据研究目标,具体实现中把各个成绩值都离散化到两个分数段:该科目成绩在前 1/3 的,记为 True,表示该课程学习效果较好,后 2/3 记为 False,表示该课程学得不好。另外,根据 Weka 的 Apriori 模型的要求,属性值 False 应该用 NULL 表示。在分析中,NULL 表示该属性没有出现。

预处理之后得到的挖掘数据源格式如图 3-2 所示,可从本书提供的 SQL Server 数据文件 scoreMining.mdf 获得,使用 SQL Server 的数据库附加功能可以重建该数据库,其中只有一个数据表 scoreMining。

eng	math	sports	political	phisical	philosophy	introduction	algoriths	discrete	datastructure	electronic	os	DB	h...	software
NULL	NULL	True	NULL	NULL	NULL	True	True	NULL	NULL	NULL	NULL	True	NULL	True
NULL	True	NULL	NULL	NULL	NULL	NULL	True	NULL	True	NULL	NULL	NULL	NULL	NULL
True	True	NULL	NULL	NULL	True	True	True	NULL	NULL	NULL	NULL	NULL	NULL	NULL
True	True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	NULL	NULL	NULL	NULL	NULL
True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	NULL	True	NULL	NULL
NULL	NULL	NULL	NULL	NULL	True	NULL	NULL	NULL	NULL	NULL	True	True	NULL	NULL
NULL	NULL	NULL	True	True	True	NULL	True	True	True	True	True	True	True	True
NULL	NULL	True	NULL	True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
NULL	NULL	True	NULL	NULL	True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	True
NULL	NULL	NULL	True	NULL	NULL	NULL	True	True	True	True	True	True	True	True
NULL	NULL	True	True	NULL	True	True	True	NULL	True	True	NULL	True	NULL	True
NULL	NULL	NULL	NULL	NULL	True	True	NULL	True	NULL	NULL	NULL	NULL	NULL	NULL
NULL	True	True	NULL	True	NULL	NULL	NULL	NULL	NULL	True	True	NULL	True	NULL
NULL	NULL	NULL	True	True	True	True	NULL	NULL	True	True	True	True	NULL	True
NULL	NULL	NULL	NULL	NULL	True	True	True	NULL	True	NULL	NULL	NULL	NULL	NULL
True	NULL	True	True	NULL	NULL	True	NULL	True	True	NULL	NULL	True	NULL	NULL
True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	True
True	NULL	NULL	True	True	True	NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	NULL
NULL	NULL	NULL	True	NULL	True	True	True	NULL	True	True	NULL	True	NULL	True
NULL	True	NULL	True	True	True	True	True	NULL	NULL	True	True	True	True	True
NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	NULL	True	NULL	NULL	NULL	NULL	NULL
NULL	NULL	NULL	NULL	NULL	NULL	NULL	True	True	NULL	True	NULL	True	NULL	True

图 3-2 数据离散化结果

6. 生成挖掘数据文件

根据附录 B 所述的数据转换方法将数据从 SQL Server 导出为 CSV 文件,然后再转换为 ARFF 文件,作为挖掘数据源。

3.3.2 建立和使用知识流

KnowledgeFlow 模块是一个图形界面环境,除了具有 Explorer 的所有功能之外,还提供一些 Explorer 不具有的功能,如增量处理。用户可以从每个工具条中选择需要的组件放置在画

布上，并且把他们连接起来形成一个处理和分析数据的“知识流”，在 KnowledgeFlow 中可以选择使用 Weka 中提供的所有核心组件。

在 KnowledgeFlow 模块中，可以对数据进行批量处理或增量处理(Explorer 模块只能对数据进行批量处理)。目前在 Weka 中集成了 5 种支持增量处理的分类器：NaiveBayesUpdateable、IB1、IBk、LWR(locally weighted regression)和 RacedIncrementalLogitBoost。

KnowledgeFlow 模块具有以下特点：

- 直观的数据流布局风格；
- 处理批量或增量数据；
- 以线程方式实现并行地处理多分支或多数据流；
- 将各种过滤器链接在一起；
- 可以观察交叉验证中分类结果的情况；
- 对增量分类器的处理过程可视化。

下面用知识流模块解决上面的问题。创建一个知识流，采用批量方式装入 ARFF 文件，采用 Apriori 算法执行数据挖掘。

(1) 运行 WEKA 的主程序，出现 GUI 后，单击进入 KnowledgeFlow 模块，KnowledgeFlow 模块的主界面如图 3-3 所示。

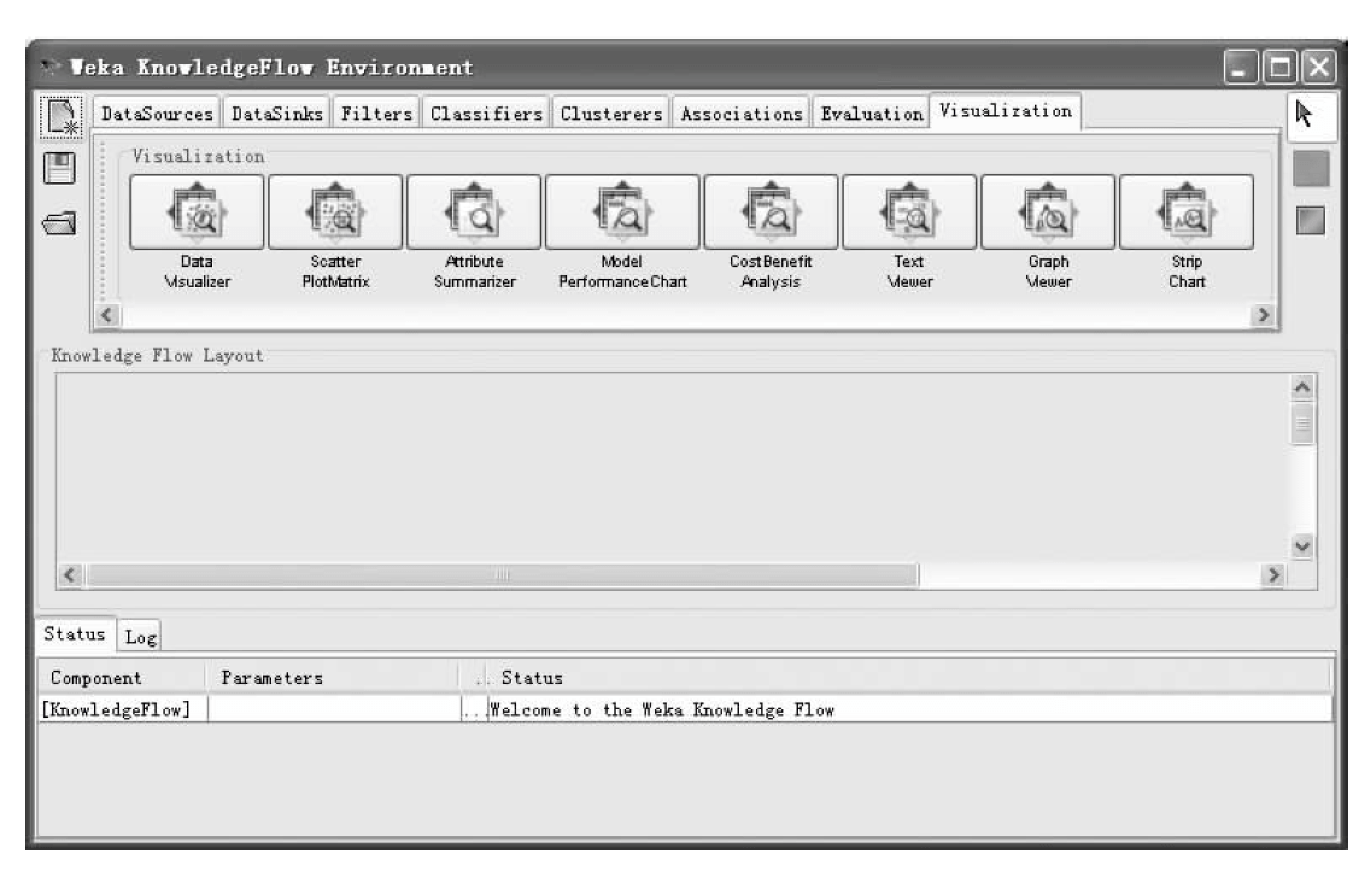


图 3-3　KnowledgeFlow 模块主界面

KnowledgeFlow 模块包含的所有组件都可在此页面上获得，它们被分类安排在相应的面板上。包括：

DataSources：包括了 Weka 的所有装载器，提供各种类型文件的装载功能。

DataSinks：包括了 Weka 的所有保存功能，可以对各种类型文件进行保存。

Filters：包括了 Weka 的所有过滤器，支持所有的数据预处理。

Classifiers：包括了 Weka 的所有分类器，供用户选择分类算法。

Clusterers：包括了 Weka 的所有聚类算法，供用户选择聚类算法。

Associations：包括了 Weka 的所有关联规则算法，供用户选择关联规则算法。

Evaluation：提供数据准备功能，包括 TrainingSetMaker，将一个数据集设置为训练集；TestSetMaker 将一个数据集设置为测试集；CrossValidationFoldMaker 按折划分数据集、训练集或测试集。TrainTestSplitMaker 把一个数据集划分成训练集和测试集；ClassAssigner 为数据集、训练集或测试集分配类属性；ClassValuePicker 规定“正类”的取值；ClassifierPerformanceEvaluator 评估批量分类器的性能；IncrementalClassifierEvaluator 评估增量分类器的性能；ClustererPerformanceEvaluator 评估批量聚类的性能；PredictionAppender 为一个测试数据集添加分类预测。

Visualization：功能与 Explorer 的可视化相同。DataVisualizer 以单一的二维散点图形式显示可视化数据；ScatterPlotMatrix 用多个小的散点图形成的矩阵显示可视化数据；AttributeSummarizer 以矩阵形式显示属性直方图，每个直方图代表一个属性特征；ModelPerformanceChart 显示可视化的阈值曲线；TextViewer 显示文本数据，如文本形式的数据集或分类性能统计信息等；GraphViewer 显示可视化的树形模型；StripChart 滚动显示数据，用于观察增量分类器的执行性能。

（2）单击 DataSources 标签进入数据源选择面板，从工具栏中选择 ArffLoader 项，此时鼠标的光标将变成一个十字形状。

在 Knowledge Flow Layout 区域的任何位置单击，就会出现一个 ArffLoader 图标，可以用鼠标拖曳这个图标，把它摆放在合适的位置，如图 3-4 所示。

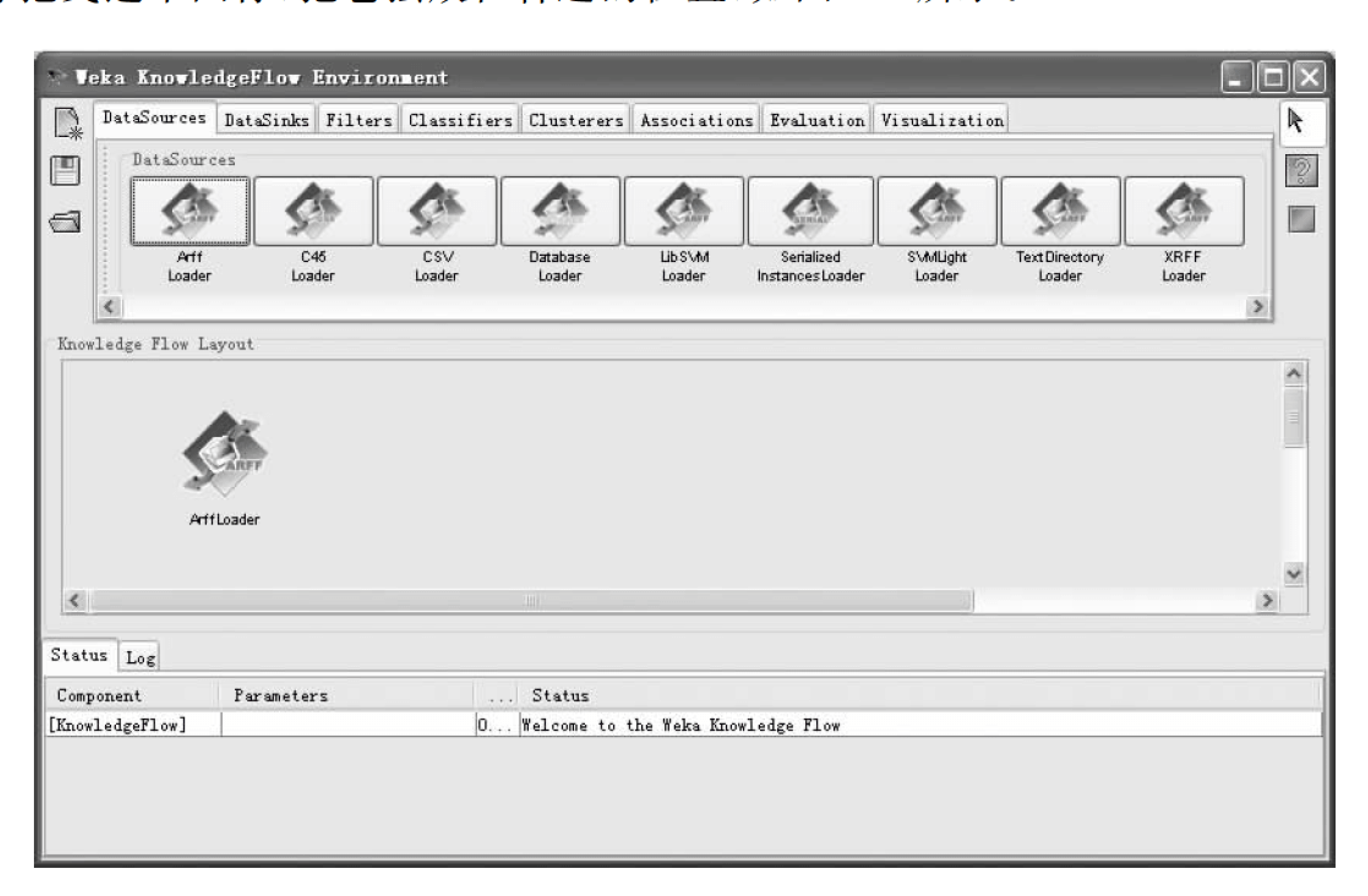


图 3-4　拖曳 ArffLoader 图标

右击此数据源图标，将会弹出的快捷菜单，选择 Configure 菜单项，打开查找文件对话框，选择要装入的 ARFF 文件，如 scoreMining. arff。也可以在数据源图标上双击，打开查找文件对话框，选择要装入的 ARFF 文件，此处选择 scoreMining. arff。

（3）选择 Associations 标签打开分类器面板，在工具栏的左边找到 Apriori 图标，并将其放置在布局中。

右击此数据源图标，从弹出的快捷菜单中选择 Data Set 菜单项，此时将会出现一条从 ArffLoader 图标出发的红色箭头，用鼠标拖曳箭头，使其指向 Apriori 图标，然后单击“确

定”按钮，即将 Apriori 图标的数据源设置为 ArffLoader 图标所装入的数据集。

在默认情况下，挖掘结果只显示最前面的 10 条规则，如果想看到更多规则，可以对默认属性设置进行修改。方法是使用 Apriori 图标的右键菜单，选择“configure”菜单项，打开参数配置对话框，修改 numRules 的值，如 20，则能够看到更多的挖掘结果。

(4) 从 Visualization 标签进入可视化面板，向布局中放置一个 TextViewer 组件，从 Apriori 图标的右键菜单选择 text 选项，并指向 TextViewer 图标，将这两个组件联系起来。

(5) 执行知识流。从 ArffLoader 图标的右键菜单选择 Start loading 选项，可以看到布局中某些图标开始显示动画效果，并且可以在界面底部的状态栏和 Log 中看到处理信息。根据数据集的大小不同和数据密集程度不同，执行时间也不同。以上步骤完成之后的结果如图 3-5 所示。

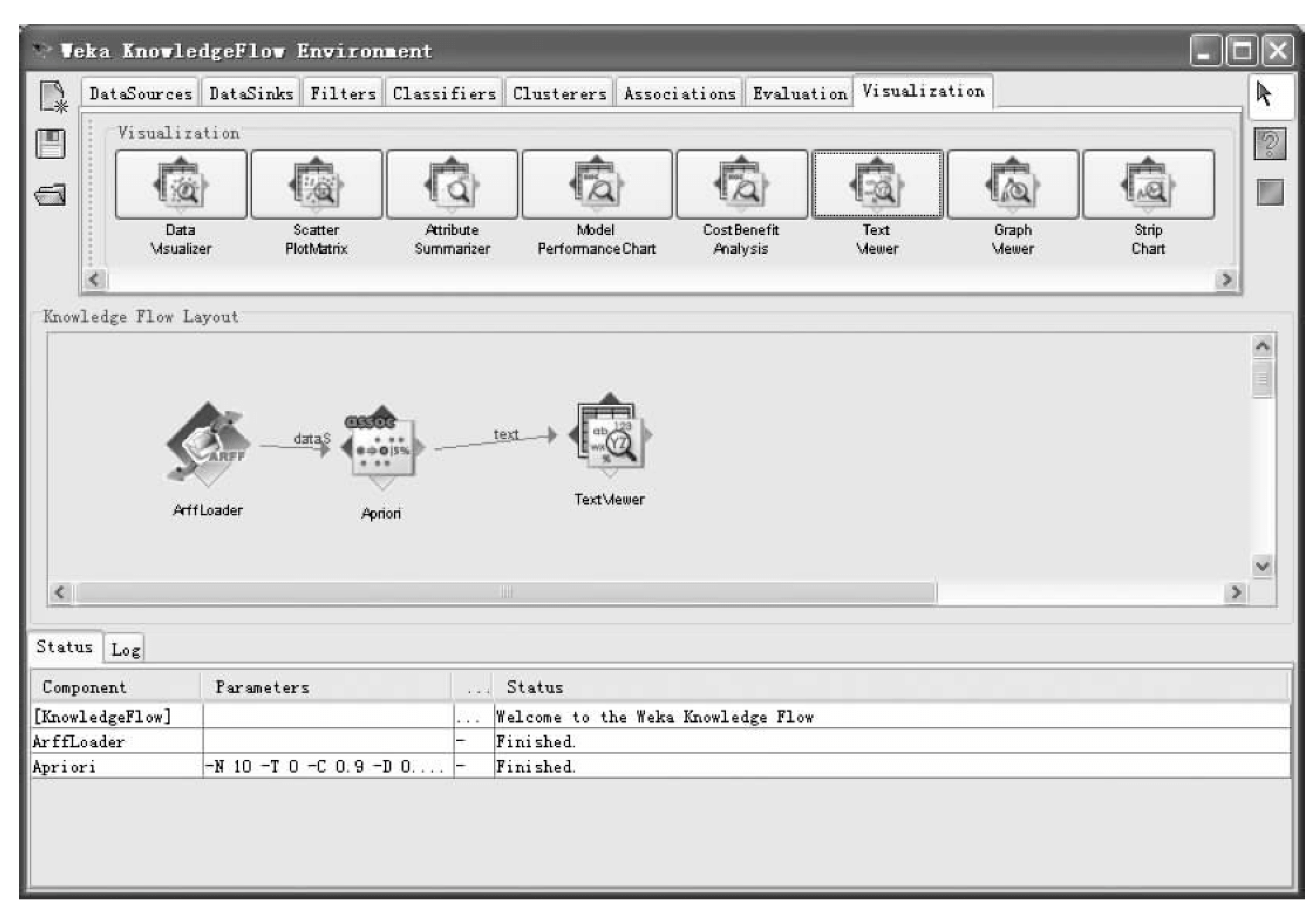


图 3-5　处理完成

(6) 查看处理结果。可以通过 TextViewer 组件的 Show Results 菜单实现，如图 3-6 所示。

其中显示了所用算法、配置参数、所用数据集，以及处理的各种性能数据。因为参数设置中默认显示前 10 条规则，所以这里只有 10 条结果。

规则以以下形式输出：

```
规则编号.规则前项 支持度==>规则后项 支持度 置信度
```

根据课程开设的前后关系，应该是根据公共课成绩预测专业基础课成绩等，根据基础课成绩预测专业方向，所以在结果中，更关心蕴含式右项为 hardware 或 software 的规则，经过筛选，得到一些规则。

```
Text Viewer
Result list
15:54:54 - Model: Apriori
Text
=== Associator model ===

Scheme:   Apriori
Relation: scoremining1

Apriori
=======

Minimum support: 0.1 (17 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 104

Size of set of large itemsets L(3): 168

Size of set of large itemsets L(4): 67

Size of set of large itemsets L(5): 1

Best rules found:

 1. phisical=True discrete=True hardware=True 19 ==> os=True 19    conf:(1)
 2. discrete=True datastructure=True hardware=True 19 ==> os=True 19    conf:(1)
 3. discrete=True datastructure=True software=True 17 ==> os=True 17    conf:(1)
 4. phisical=True discrete=True datastructure=True hardware=True 17 ==> os=True 17    conf:(1)
 5. phisical=True introduction=True software=True 20 ==> electronic=True 19    conf:(0.95)
 6. political=True phisical=True introduction=True 19 ==> electronic=True 18    conf:(0.95)
 7. philosophy=True os=True hardware=True 19 ==> political=True 18    conf:(0.95)
 8. phisical=True introduction=True os=True 19 ==> electronic=True 18    conf:(0.95)
 9. phisical=True datastructure=True hardware=True 19 ==> os=True 18    conf:(0.95)
10. phisical=True datastructure=True software=True 19 ==> os=True 18    conf:(0.95)
```

图 3-6　结果显示

```
6. political=True phisical=True introduction=True 19==>electronic=True 18  conf:(0.95)
16. political=True datastructure=True electronic=True 21==>DB=True 19  conf: (0.9)
17 phisical=True introduction=True electronic= True 21==>software= True 19  conf:(0.9)
20. phisical=True discrete=True datastructure=True 21==>os=True 19  conf: (0.9)
22 phisical=True discrete=True os=True 21==>hardware=True 19  conf: (0.9)
23 discrete=True datastructure=True os=True 21==>hardware=True 19  conf: (0.9)
26. discrete=True datastructure=True electronic=True 20==>os=True 18  conf: (0.9)
```

其中规则 17 所表达的是当物理、计算机导论和电工学成绩优秀时，软件方向的课程会得到好成绩；规则 22 所表达的是当物理、离散数学、操作系统成绩优秀时，硬件方向的课程会得到好成绩；规则 23 所表达的是当离散数学、数据结构、操作系统成绩优秀时，硬件方向的课程会得到好成绩。

这些规则的可用性，还要依据数据预处理的方法、离散化的取值等等有所不同。读者可以对本书提供的原始数据用不同的预处理方法进行处理，观察挖掘结果的不同。

操作完毕，直接点击右上角的图标“X”关闭窗口即可。

3.4　案例小结

在本实例中，首先根据数据特点和挖掘模型要求对获得的电子数据进行了预处理，然后选择 Weka 的 KnowledgeFlow 模块建立了一个知识流，其中挖掘模型选择了关联规则分析

的 Apriori 模型,产生了挖掘结果,并根据挖掘目标对挖掘结果进行了筛选和解读,完整演示了数据挖掘的全部过程。当然,还应该进行调查分析确定规则的可用性,这项工作超出了本书的范围,在此不再进行讨论。

读者可以从本章获得的知识有两个方面,一是了解关联规则分析的过程;二是学习 Weka 的 KnowledgeFlow 模块的使用方法。

实例 4　基于决策树方法的网球运动天气状况评价分析

4.1　任 务 描 述

搜集到 14 天的天气情况如表 4-1 所示，用 outlook，temperature，humidity 和 wind 四个条件属性来描述。其中，属性 outlook 为离散型属性，取值分别为 sunny、rainy、overcast；属性 windy 为离散型属性，取值分别为 FALSE、TRUE；属性 temperature 和 humidity 为数值型属性。最后一个属性 play 为决策属性，属性值为 yes 是说这天适合打网球，属性值为 no 是说这天不适合打网球。

表 4-1　天气情况数据

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

（1）试根据这些数据建立评价是否适合打网球的评价规则。

（2）若给定一天的天气为 overcast，66.0，78，FALSE，问这天是否适合打网球？

4.2　技 术 原 理

4.2.1　决策树的概念

决策树方法是最受欢迎的数据挖掘技术之一，主要用于分类和预测。决策树学习是以样本为基础的归纳学习方法，将决策树转换成分类规则比较容易。决策树的表现形式类似于流程图的树结构，在决策树的内部结点进行属性测试，并根据属性值判断由该结点引出的分支，在叶结点得到结论。图 4-1 所示为一个决策树实例。

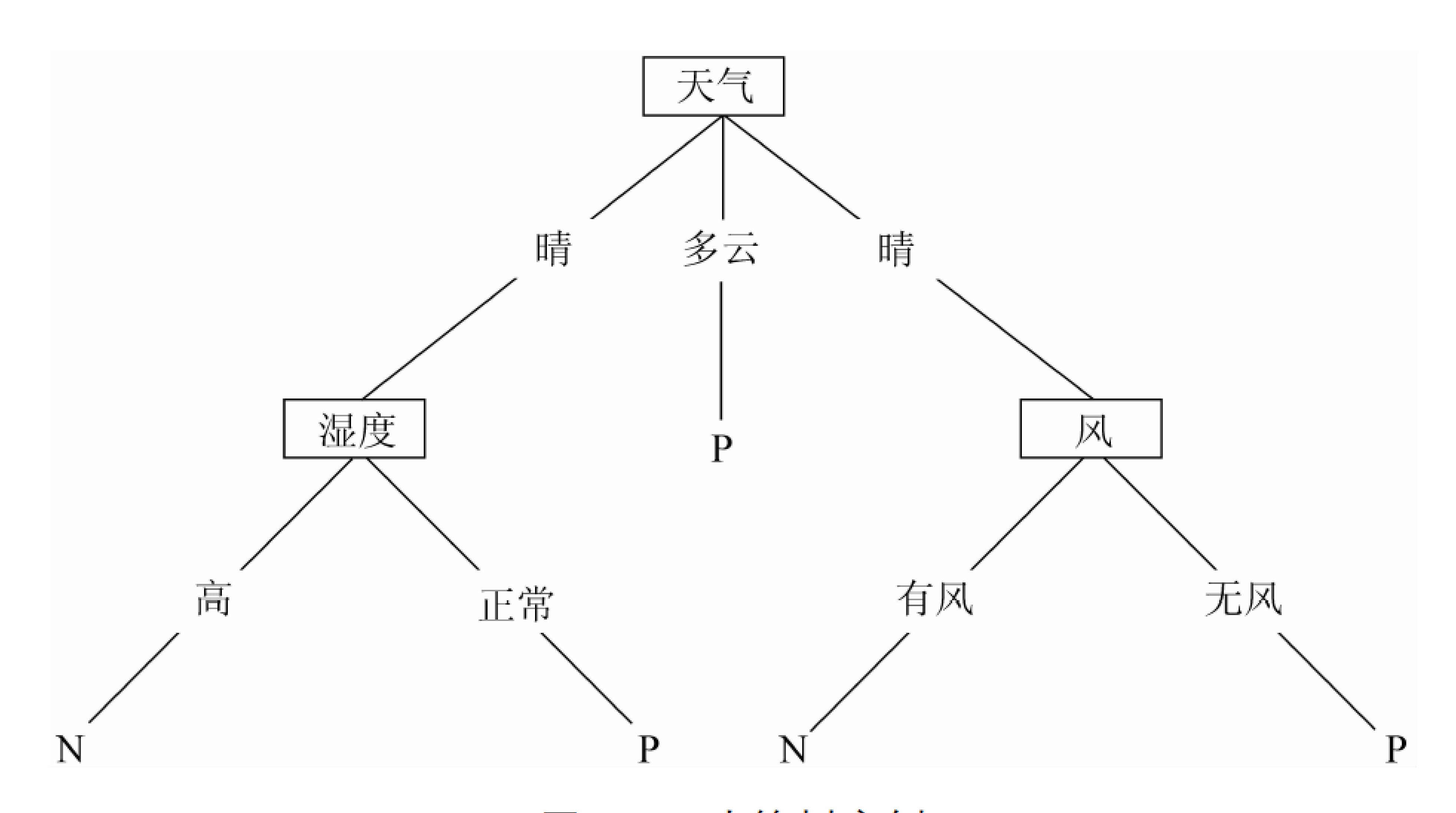


图 4-1　决策树实例

4.2.2 信息论的基本概念

消息(符号) $U_i(i=1,2,\cdots,q)$ 的发生概率 $P(U_i)$ 组成信源数学模型(样本空间或概率空间),即

$$[U,P]=\begin{bmatrix} U_1 & U_2 & \cdots & U_q \\ P(U_1) & P(U_2) & \cdots & P(U_q) \end{bmatrix} \tag{4-1}$$

信息熵用来衡量每个消息所提供的平均信息量,定义为:

$$H(U)=\sum_i P(U_i)\log_k \frac{1}{P(U_i)}=-\sum_i P(U_i)\log_k P(U_i) \tag{4-2}$$

后验熵是当信道接收端接收到输出符号 $V=V_j$ 后,关于输入符号 U_i 的信息度量,即

$$H(U\mid V)=\sum_j P(V_j)\sum_j P(U_i\mid V_j)\log \frac{1}{P(U_i\mid V_j)} \tag{4-3}$$

4.2.3 ID3 建树算法

(1) 对当前例子集合,计算各特征的互信息。

(2) 选择互信息最大的特征 A_k。

(3) 把在 A_k 处取值相同的例子归于同一子集,A_k 取几个值就得几个子集。

(4) 对既含正例又含反例的子集,递归调用建树算法。

(5) 若子集仅含正例或反例,对应分枝上标 P 或 N,返回调用处。

4.3 具体实现

具体说明:

(1) 选择"开始"→"所有程序"→Weka3.6.5→Weka3.6 选项,如图 4-2 所示。



图 4-2 打开 Weka 软件

(2) 在打开的文件中,单击 Explorer 按钮,如图 4-3 所示。

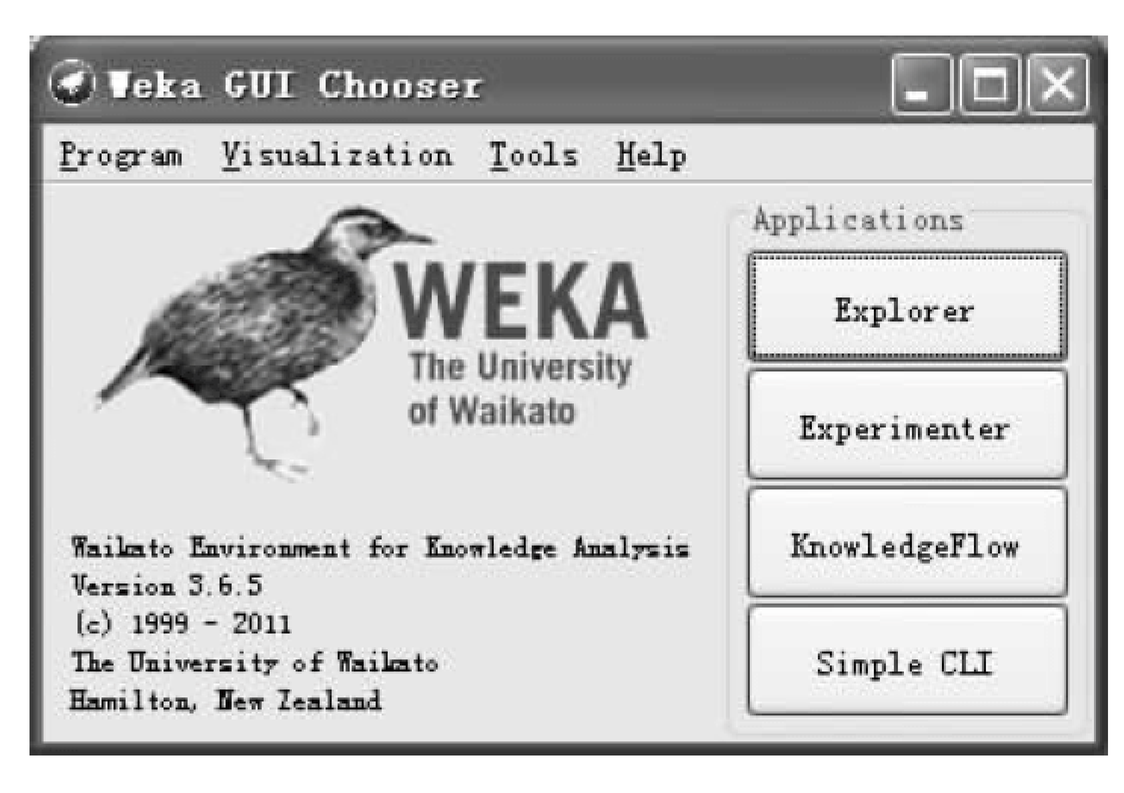


图 4-3 打开 Explorer 应用

（3）单击 Open file 按钮，选择要打开的文件 weather. arff，并单击“打开”按钮，如图 4-4 所示。

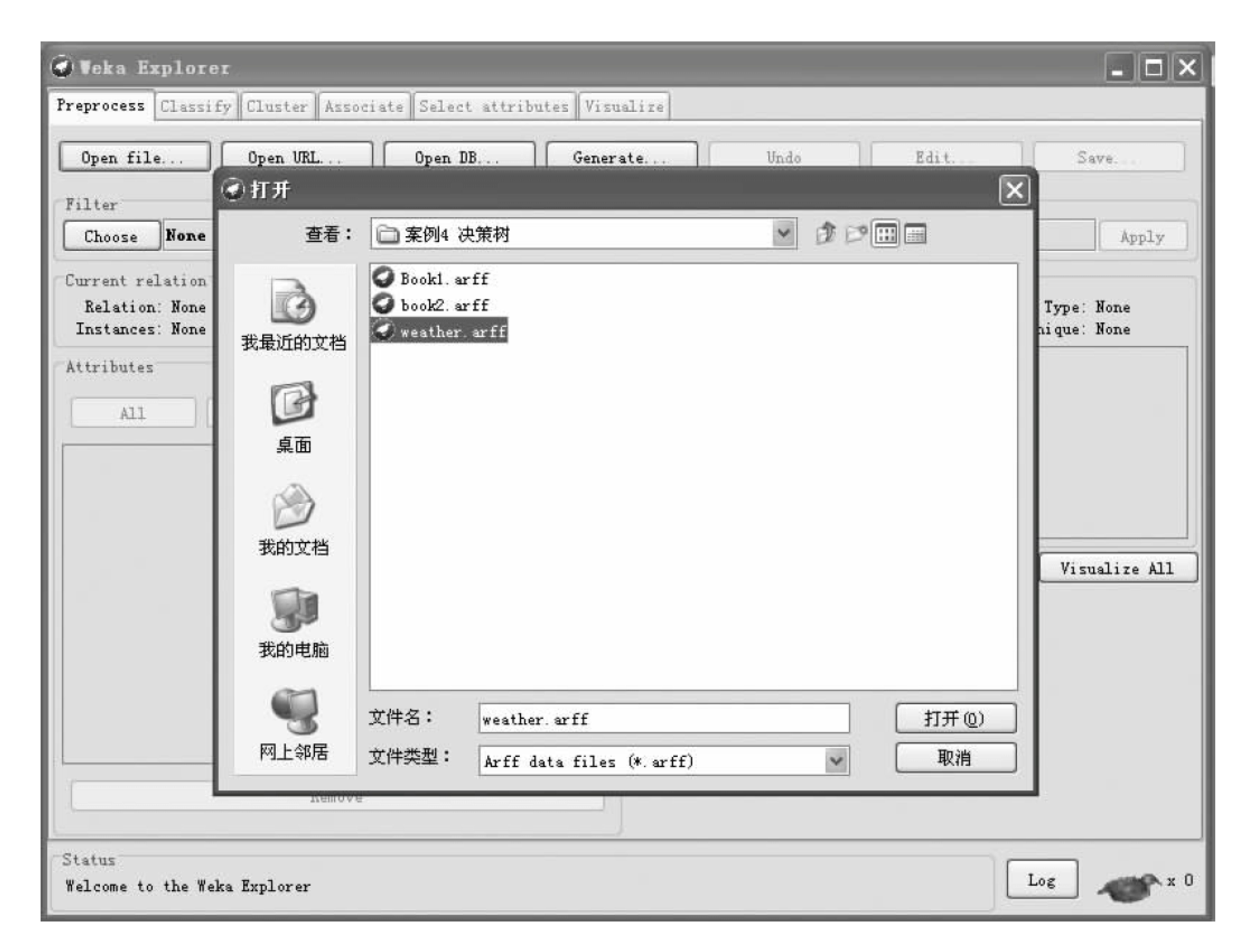


图 4-4　打开数据文件

（4）在如图 4-5 所示的界面中，可以知道 weather 数据集中共有 14 个实例，每个实例有 5 个属性。选中某个属性，可以查看 14 个实例关于这个属性的属性值的最小值、最大值、均值和标准差等信息。然后单击 Classify 标签栏，并单击 Choose 按钮。

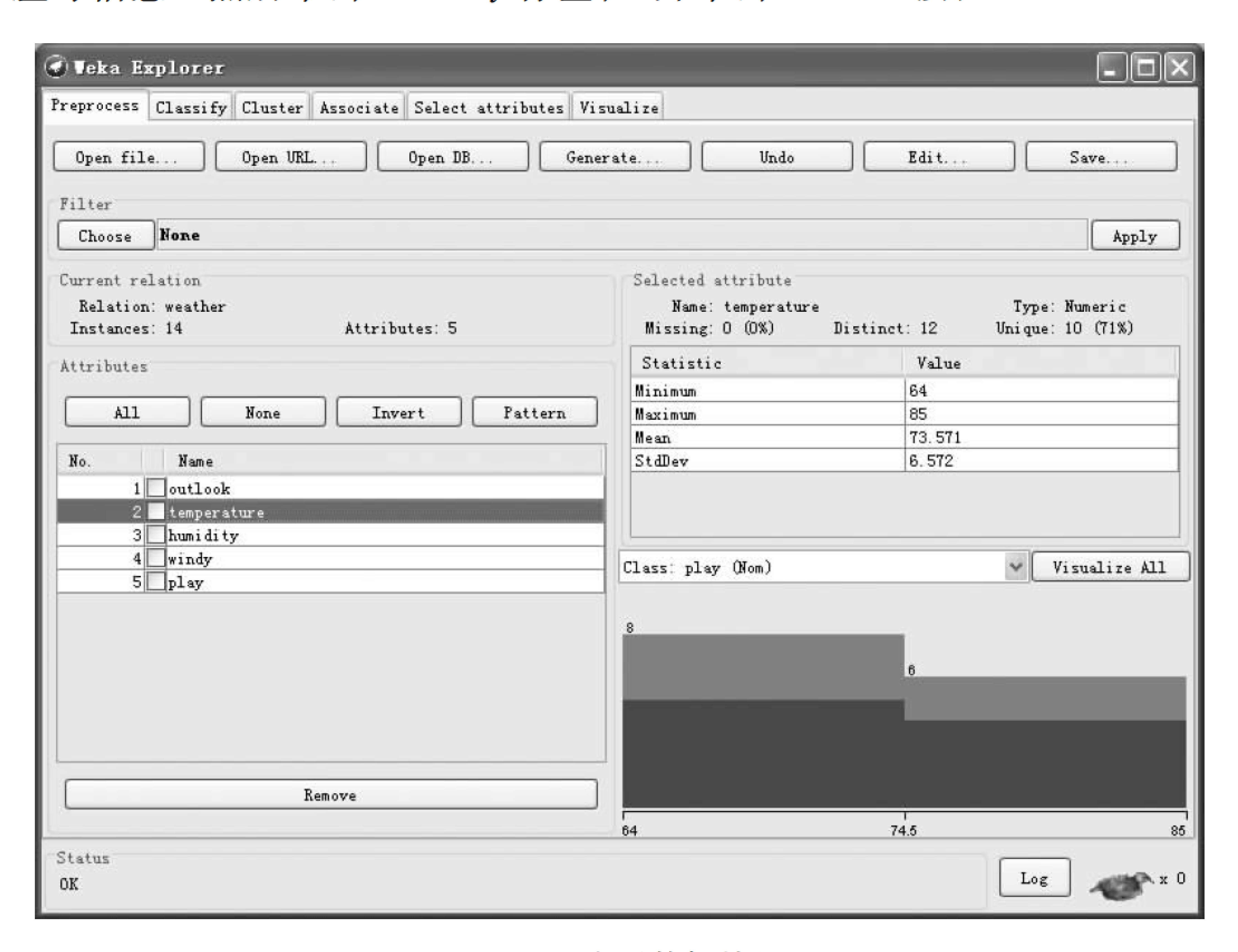


图 4-5　查看数据特征

（5）如图 4-6 所示，选择 J48 分类器，并单击 Close 按钮。

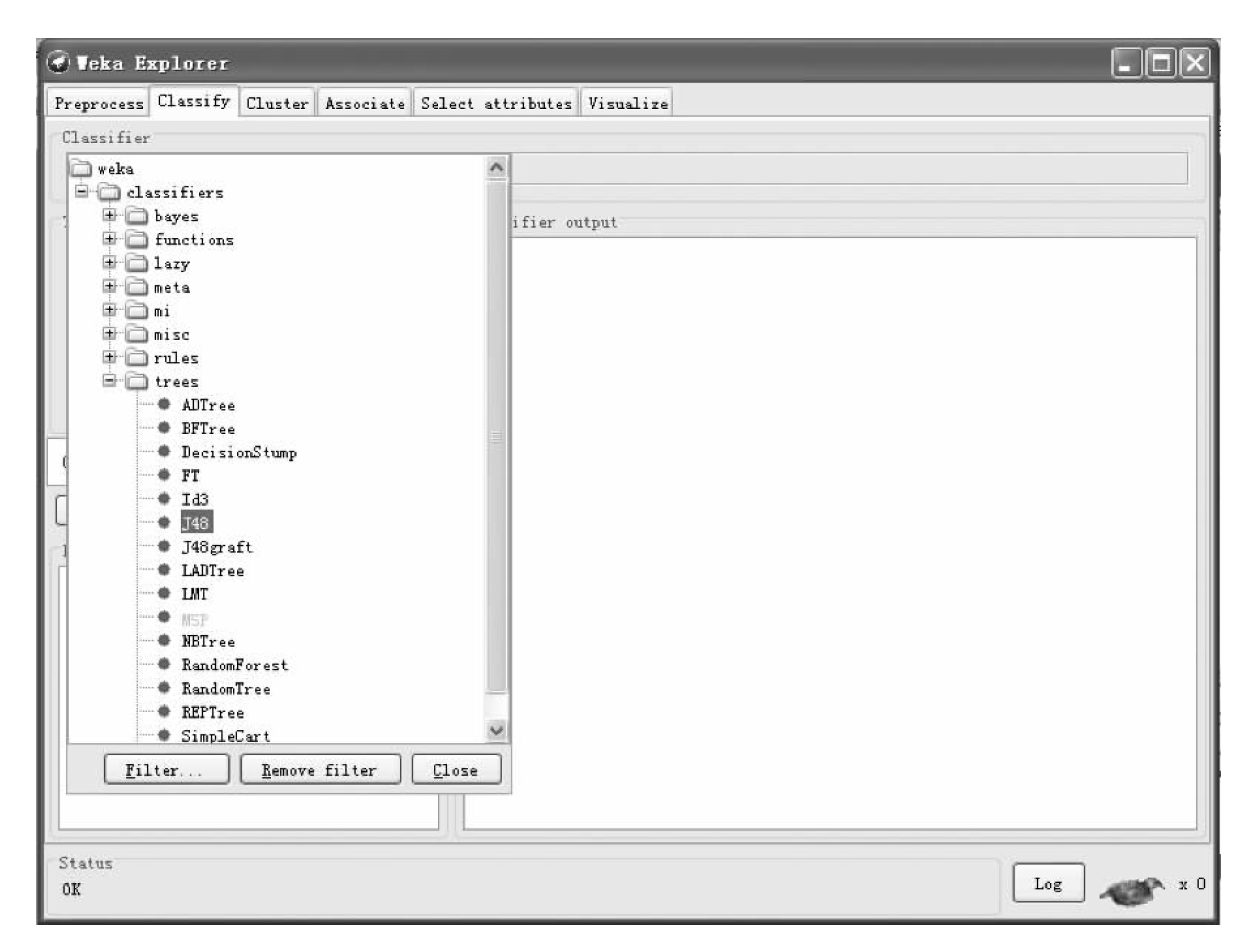


图 4-6　选择决策树方法

（6）如图 4-7 所示，建立 Book1. arff 文件。



图 4-7　建立测试文件

（7）选中 Test options 选项中的 Supplied test set 单选按钮，单击 Set 按钮，如图 4-8 所示。

（8）选中 Book1. arff 文件，并单击“打开”按钮，如图 4-9 所示。

（9）单击 Start 按钮，Weka 软件显示运行结果，如图 4-10 所示。

图 4-8　设置测试选项



图 4-9　选定测试文件

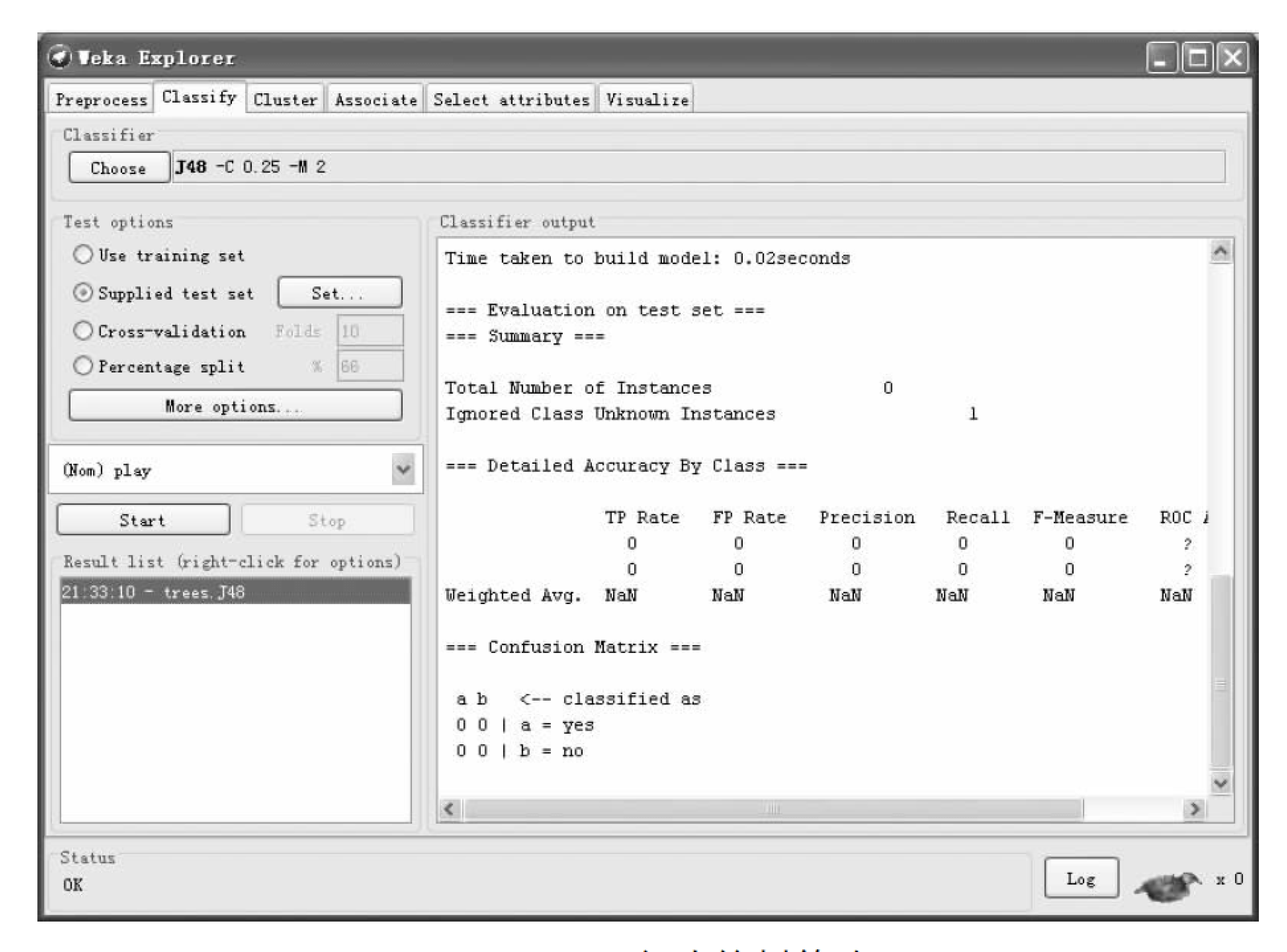


图 4-10　运行决策树算法

（10）右击 Result list 中刚才出现的那一项，在弹出的菜单中选择 Visualize tree 项，如图 4-11 所示。

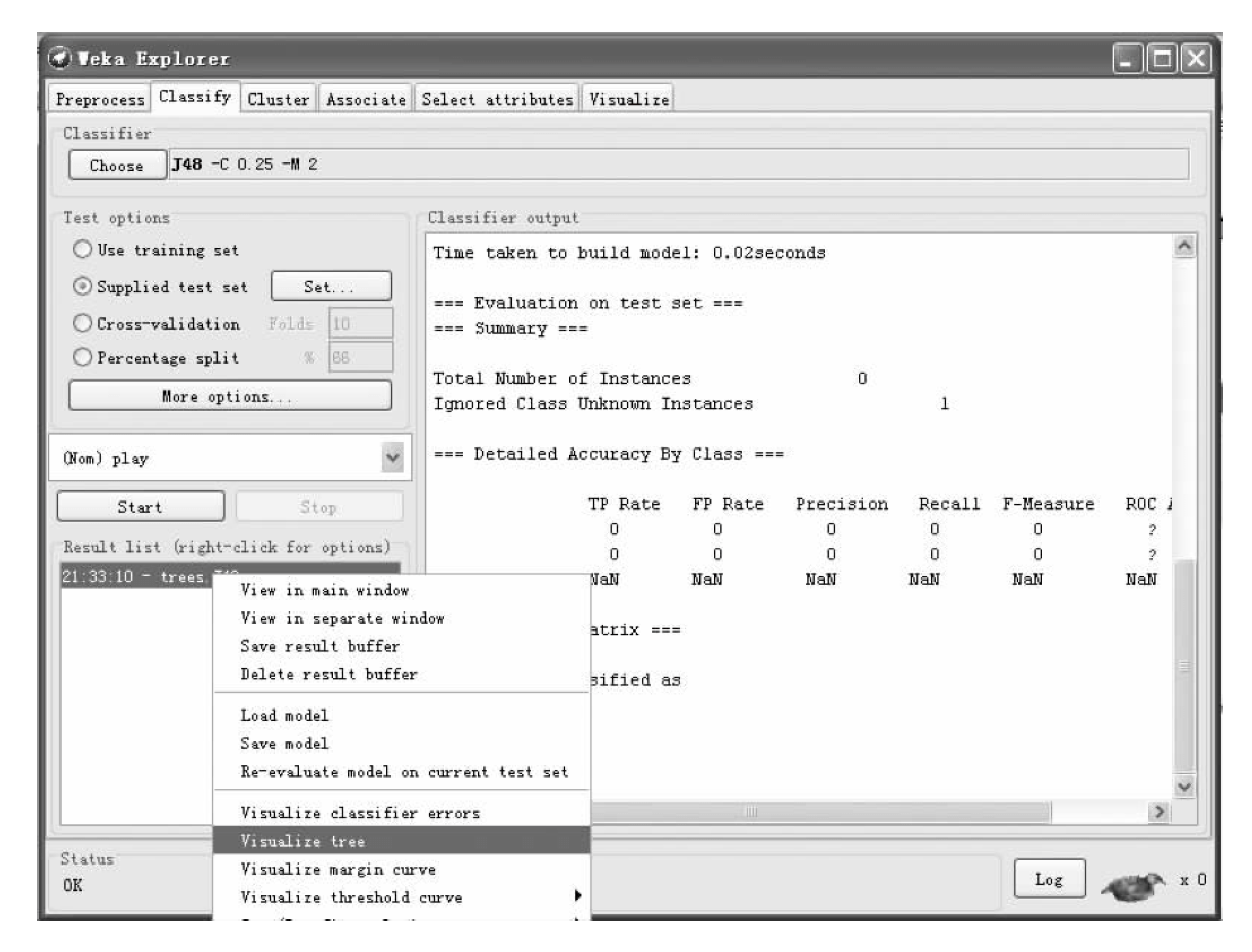


图 4-11　选择可视化决策树选项

（11）在新窗口中，可以看到图形模式的决策树，如图 4-12 所示。

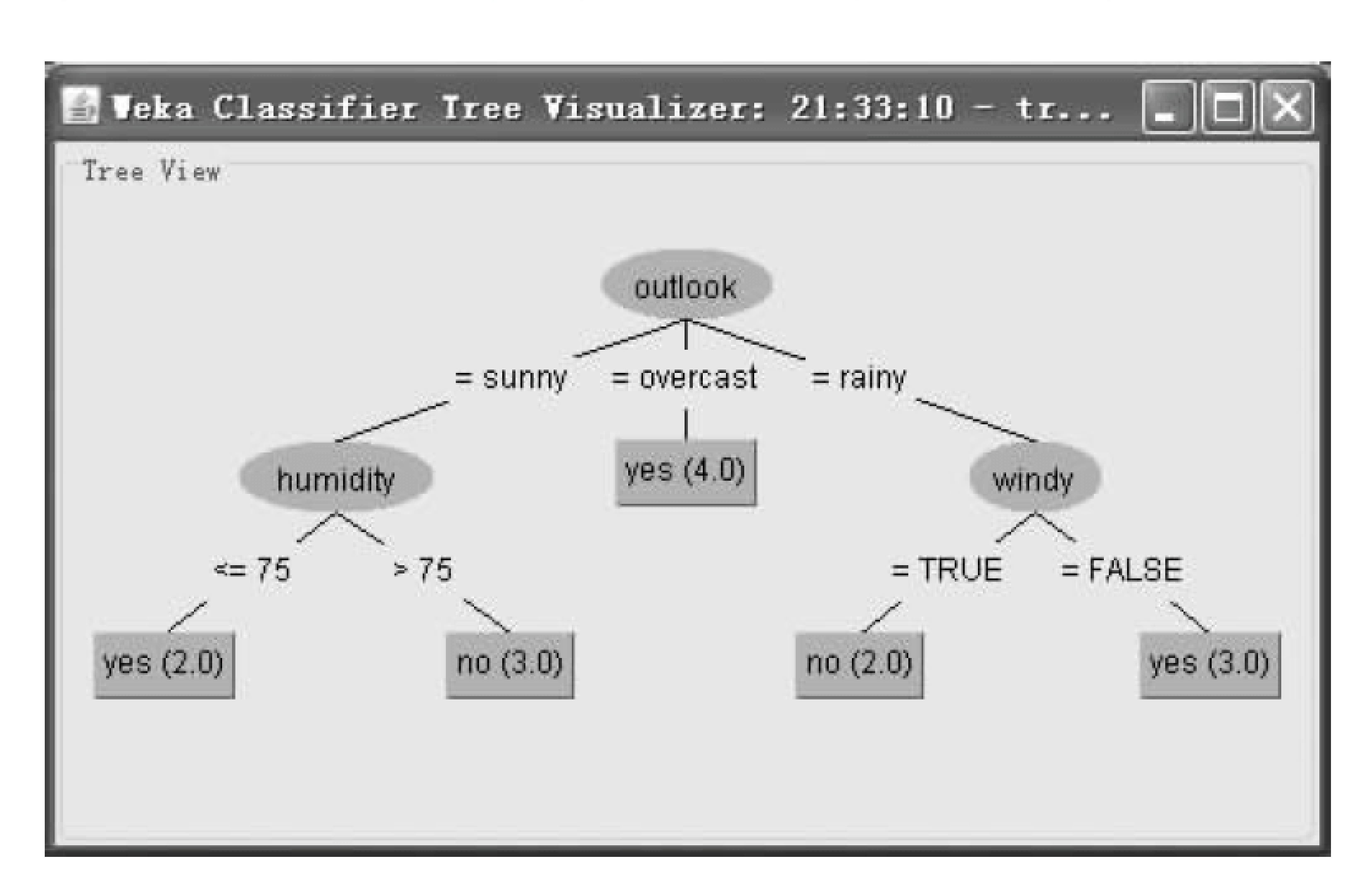


图 4-12　查看决策树

（12）右击 Result list 中刚才出现的那一项，在弹出的菜单中选择 Visualize classifier error 项，如图 4-13 所示。

（13）在弹出的对话框中，单击 Save 按钮，并保存为文件 Book2. arff，如图 4-14 和图 4-15 所示。

（14）打开文件 Book2. arff，可知若给定一天的天气为 overcast，66. 0，78，FALSE，则预测值为 yes，所以这天适合打网球，如图 4-16 所示。

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click for options)

21:38:54 - rules.ZeroR

21:40:24 - trees.J48

Classifier output

```
Time taken to build model: 0.02seconds

=== Evaluation on test set ===
=== Summary ===

Total Number of Instances                0
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
                 0         0          0          0        0          ?
                 0         0          0          0        0          ?
Weighted Avg.   NaN       NaN        NaN        NaN      NaN        NaN
```

View in main window

View in separate window

Save result buffer

Delete result buffer

Load model

Save model

Re-evaluate model on current test set

Visualize classifier errors

Visualize tree

Visualize margin curve

Status

OK

Log x 0

图 4-13 选择 Visualize classifier errors 选项

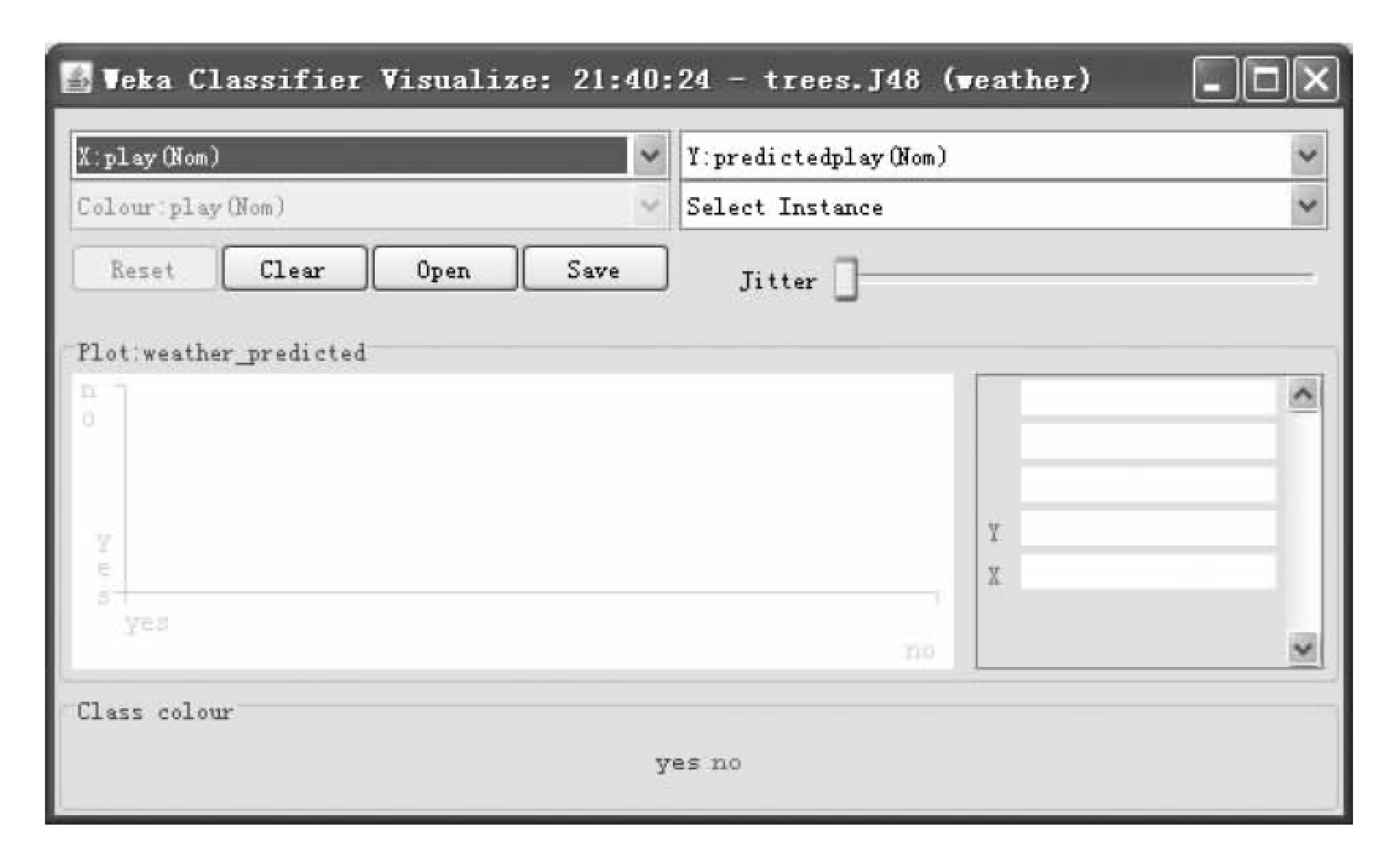


图 4-14 查看可视化分类结果



图 4-15 将分类结果保存于文件

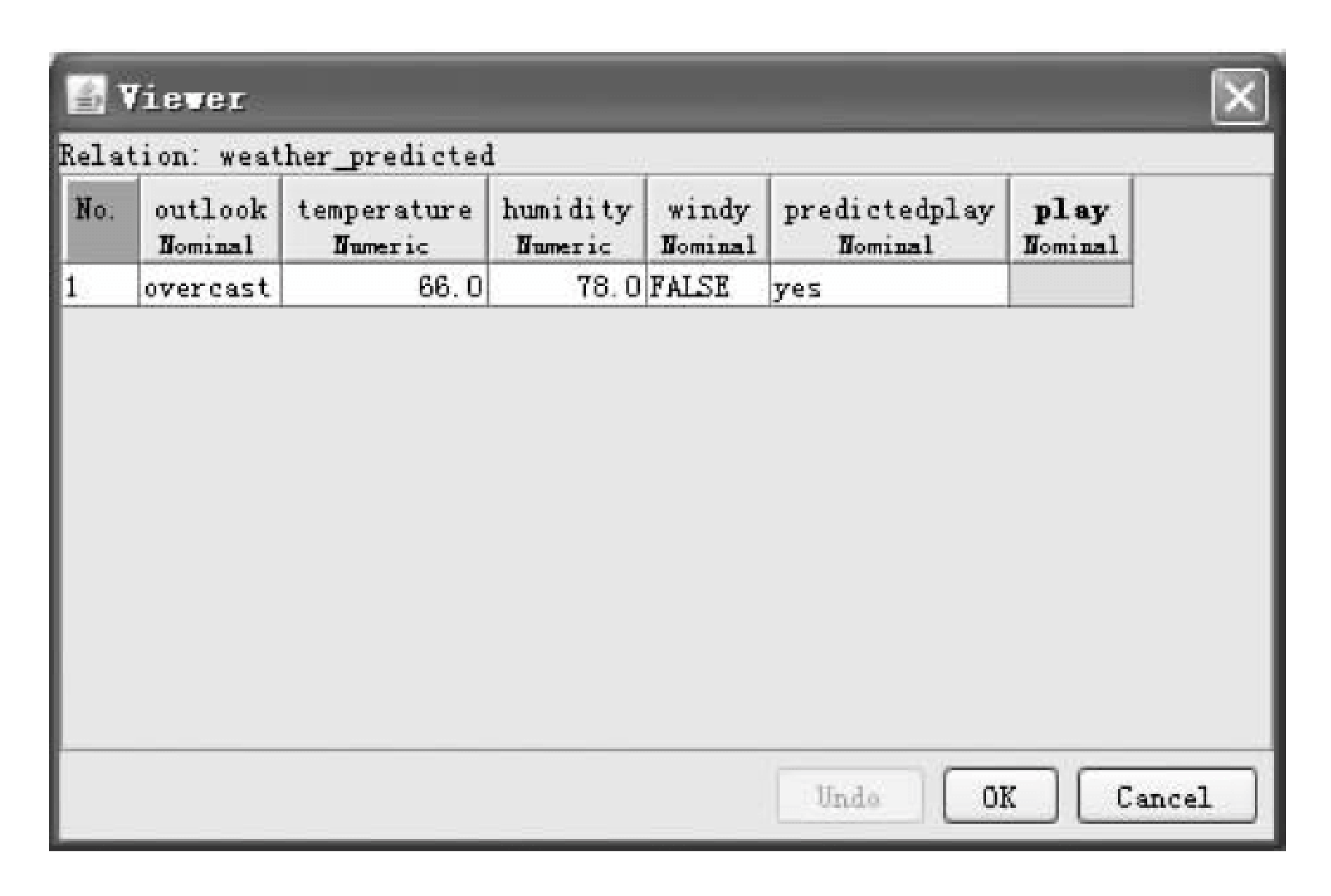


图 4-16 查看预测结果

4.4 案例小结

决策树方法是最受欢迎的数据挖掘技术之一，主要用于分类和预测。决策树学习是以样本为基础的归纳学习方法，利用信息论原理来建立决策树模型。决策树方法可以高度自动化地建立易于为用户所理解的模型，实用效果较好。本案例利用 Weka 软件自带的网球运动天气状况数据，采用分类技术中的 J48 决策树构建方法，得到了直观的决策树模型。在此基础上，利用得到的决策树模型，可以预测给定的某种天气状况是否适合进行网球运动，取得了满意的结果。

实例5　基于 Weka Experimenter 模块的人力资源管理挖掘模型选择分析

5.1　任务描述

人力资源管理的目的在于结合企事业单位发展的需要,获得企事业单位所需要的员工,并且创造条件以保证员工能完全投入工作,充分发挥他们的潜能,所以人力资源管理对企事业单位的生存发展起着至关重要的作用。

目前,中国高校的人力资源结构基本上由5个部分组成:教学科研人员、管理人员、服务人员、离退休人员和附属部门的工作人员。利用数据挖掘技术对高校人力资源数据源中的数据进行分析,寻找其中有价值的关系和规律,对管理人员职位的安排、教师聘用、培养、选拔等实际工作能够起到一定程度的辅助作用,进而提供决策支持。

解决同一个应用问题不仅可以采用不同的数据挖掘类型,如分类、聚类、关联规则等,即使确定了挖掘类型之后,也有多种挖掘模型可供选择。Weka 的 Experimenter 模块专门设计用来评估各种方法及不同参数设置情况下的挖掘结果,可以就某个数据集在多个挖掘模型之间进行比较。本实例使用人力资源数据进行数据挖掘,用 Experimenter 模块对两种候选算法进行比较,决定最终选用的挖掘模型(并不具体实现挖掘任务)。读者可以采用本书提供的数据,用 Explorer 模块或 KnowledgFlow 模块完成挖掘任务。

5.2　技术原理

5.2.1　挖掘类型确定

本问题拟根据高校职工的各项背景信息,预测他们在不同工作岗位上可能做出的成绩,其中对成绩的评价以考核成绩为参考,所以挖掘任务是对职工进行分类。本实例的主要任务是在确定了采用分类方法解决问题之后,对 Weka 的 Experimenter 模块提供的分类模型进行比较,确定哪种模型最适合此问题。

5.2.2　数据收集和准备

数据来源于不同院校的人力资源数据库,而且来自多个部门。例如,教职工基本情况数据来自人力资源部门,科研成果来自科研管理部门,教学考核来自教务部门等。数据来源复杂,在放入挖掘库之前,要进行整理,即对数据进行预处理。

用预处理过的数据建成挖掘库,数据挖掘库中的数据不同于原始数据,是符合数据挖掘要求的数据。在本实例中,数据源是经过整理的某高等院校教职工人力资源数据库。

5.3 具体实现

5.3.1 数据预处理

1. 数据集成

获得的样本数据有的存储在 Microsoft SQL Server 2000 数据库中，还有的存储在 Microsoft Access 2000 数据库中，需要将这些数据集成到一起，并且要使那些本来存在冲突和不一致的数据一致化。不同的数据库的数据定义通常都存在很大的差异，如同样表示教职工的年龄信息，有的可能使用整数表示实际年龄，有的可能使用出生年月。这都需要使用数据集成的相关方法进行处理。

2. 选择数据

应该去掉那些肯定和挖掘无关的数据，如姓名，只保留本次数据挖掘所需要的数据。注意，必须保留类似“职工编号”这样的主键信息。

3. 数据清理

由于各种各样的数据质量问题，数据中可能包含了不正确的值、空缺值。而且从多个不同的源集成数据时，不同数据源之间的数据存在不一致。

在人力资源数据库中，空缺值除了因录入员操作失误没有输入以外，一般都代表“无”，如无职务或无党派等。对于操作失误导致的空缺值，通过各字段间关系的推断，或是询问数据来源单位核实可以填充完整。对于代表“无”的空缺值，可以用特定的值来代替。

对于不一致数据，可以通过人工纠正的方法处理。

4. 数据离散化

对于给定的数值属性，可以通过概念分层来进行离散化，概念分层通过用较高层的概念（如年龄的老、中、青）替换较低层的概念（如年龄的具体数值）来达到归约数据的目的。

由于人力资源库中的属性大多具有有限个不同值，可以生成分类属性的概念分层。对这些属性分层代码的确定如下：

- 职务级别：无 0 、副科 1、正科 2、副处 3、正处 4、副局 5、正局 6。
- 最高学历：初中 00 、高中 11、中技（中专）01、大专 02、学士 03、双学位 33、硕士 04、博士 05、博士后 06。
- 职称级别：无 0、初级 1、中级 2、副高 3、正高 4。
- 身份级别：行政 1、工勤及其他 2、教辅 3、教师 4、科研 5。
- 政治面貌：群众 1、共青团员 2、共产党员 3、民主党派 4。
- 性　　别：女 0、男 1。
- 专业代号：哲学（社会学政治法律）1、经济管理 2、文化教育 3、自然科学 4、农业科学 5、医药卫生 6、工程技术 7。
- 考核成绩：（95～100）A、（85～95）B、（75～85）C、（60～75）D、60 以下 E。
- 毕业学校：重点院校 A、一般院校 B、进修 C、专科 D、中技（高中）E、初中 F、留学 G。
- 出生年代：例如 60，指 60～69 年出生的人。
- 本院兼职：Y 代表是兼职。

最后，得到的用于挖掘的数据集如图 5-1 所示。

表 - dbo.HumanResources | 表 - dbo.scoreMining | 摘要

DM	KH	XB	CSND	ZZMM	ZWJB	ZCJB	RZSJ	ZGXL	BYXX
101002	B	1	50	3	5	3	80	04	C
101003	B	1	60	3	5	4	80	05	A
101004	0	0	60	1	0	0	80	01	F
101005	C	1	50	3	6	4	80	04	A
101006	0	1	50	1	2	1	90	00	F
101007	0	0	50	3	4	4	80	03	B
101008	0	1	40	3	6	4	80	03	A
101009	C	1	40	3	5	4	80	03	A
101010	0	1	40	3	5	4	70	03	A
101011	0	0	60	3	0	1	80	03	C
101012	0	1	60	3	4	0	80	04	C
101013	0	1	70	3	3	1	90	03	B
101014	C	0	50	3	2	3	80	03	B
101015	C	1	50	3	5	3	90	03	A
101016	0	1	40	3	5	3	00	03	B
102001	0	0	60	1	0	0	90	01	F
102002	E	1	60	3	4	3	80	04	A
102003	D	0	70	3	1	2	90	03	B
102004	0	1	60	3	0	2	00	03	B
103001	0	1	40	3	4	0	70	02	D
103002	0	0	70	3	0	1	90	03	B
103003	0	0	50	3	3	0	70	02	C
104001	0	1	50	3	0	2	90	11	F
104002	0	1	60	1	0	0	90	01	F
104003	0	0	60	1	0	0	80	11	F

图 5-1　用于挖掘数据集

注意：Weka 不能识别汉字，所以字段名要用英文表示。

图 5-1 中的数据表来自于本书提供的 SQL Server 数据文件 humanResourse.mdf，使用 SQL Server 的数据库附加功能可以重建该数据库，其中只有一个数据表 HumanResources。

5. 生成挖掘数据源

为了方便地使用 Weka 处理本任务，需要将数据集保存为 ARFF 格式，根据本书附录 B 所述的数据转换方法将数据从 SQL Server 导出为 CSV 文件，然后再转换为 ARFF 文件，作为挖掘数据源。

5.3.2 模型比较和选择

用 Weka 的 Experimenter 模块对几种挖掘模型进行比较，从而选择适合的模型。

（1）从 Weka GUI 首页单击 Experimenter 按钮，打开 Experimenter 首页面，如图 5-2 所示。

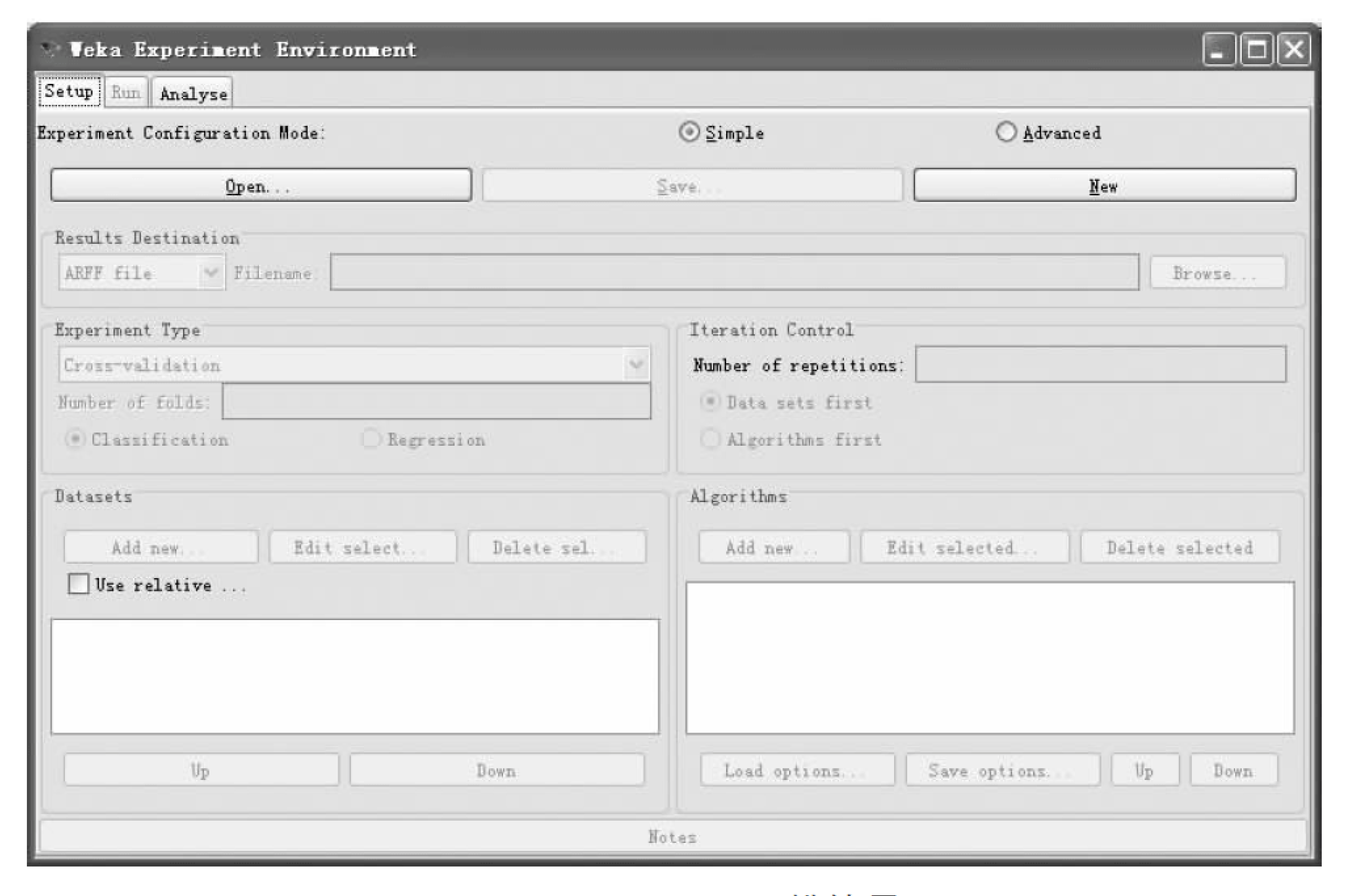


图 5-2　Experimenter 模块界面

（2）Experimenter 有两种配置模式：简单模式和高级模式。本节只使用简单模式进行模型的建立和比较。首先，要创建一个新的实验，单击 New 按钮使得各个配置区域可用，如图 5-3 所示。

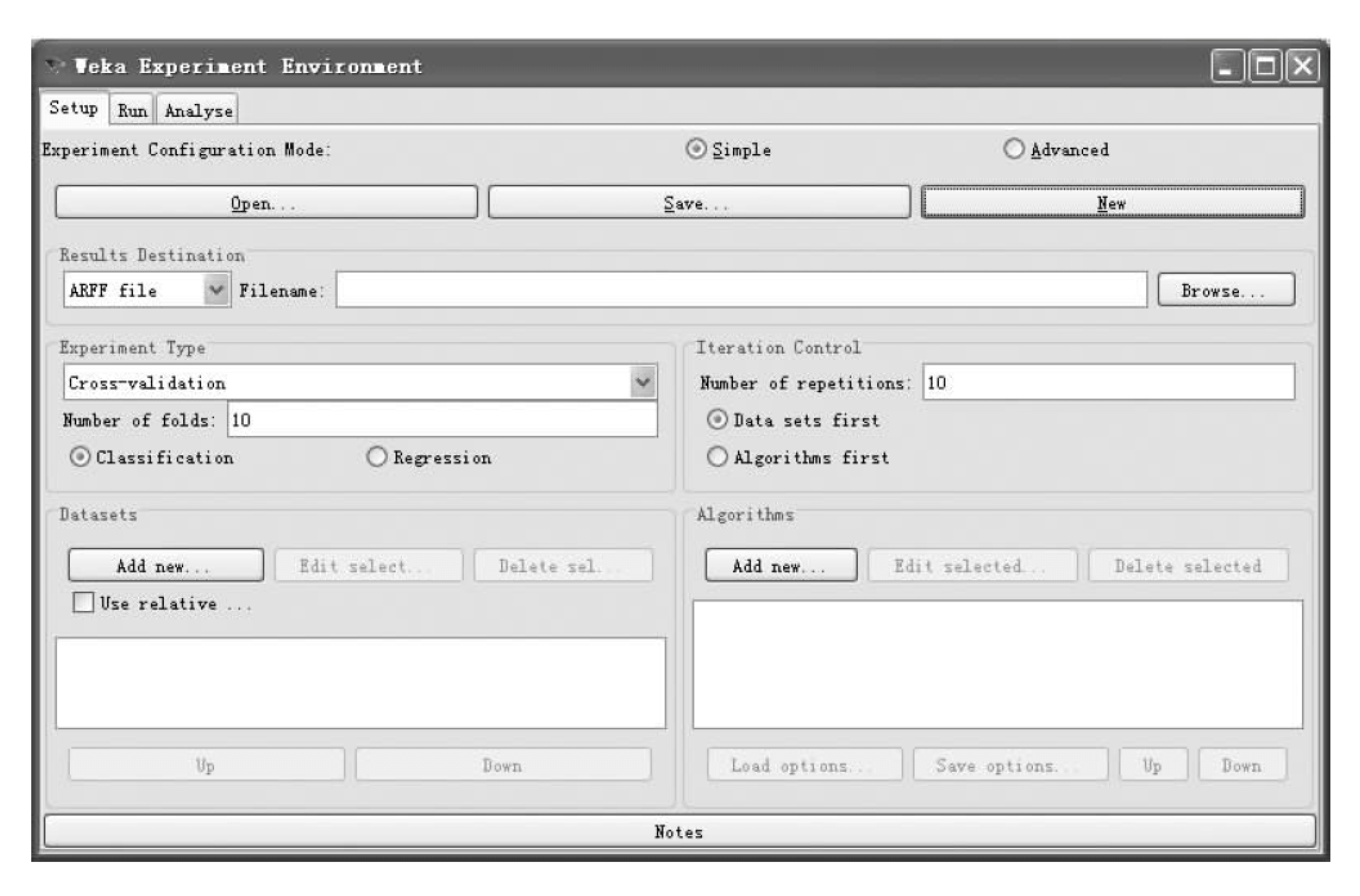


图 5-3　新建实验

页面最上方是两个配置模式选择按钮 Simple 和 Advanced。

接下来是 3 个功能按钮，分别为 Open、Save 和 New，用来打开、保存和创建新的实验。

- Result Destination 区域用来指定实验结果保存的目的文件。
- Experiment Type 区域可以指定实验类型，如“交叉验证”或“训练/测试”，同时指定交叉验证的折数，或训练/测试数据集的百分比，还可以选择实验是分类还是回归。
- Iteration Control 区域设置迭代次数，说明实验选择“数据集优先”还是“算法优先”。
- Datasets 区域用来选择数据集。
- Algorithms 区域用来选择算法。

在结果目标文件浏览框中选择结果文件类型，可选项包括 ARFF 文件、CSV 文件和 JDBC Database 文件。并通过浏览按钮选择或创建一个目标文件，实验类型选择为交叉验证，折数 10。保持迭代控制参数不变。

（3）单击 Datasets 区域的 Add new 按钮添加数据集，选择上面转换好的“E：/书稿/实例/人力资源/humanResourse. arff”数据集。单击 Algorithms 区域的 Add new 按钮选择算法，如图 5-4 所示。



图 5-4　添加数据集

（4）在图 5-4 中的对话框中，单击 Choose 按钮，选择算法，如图 5-5 所示。

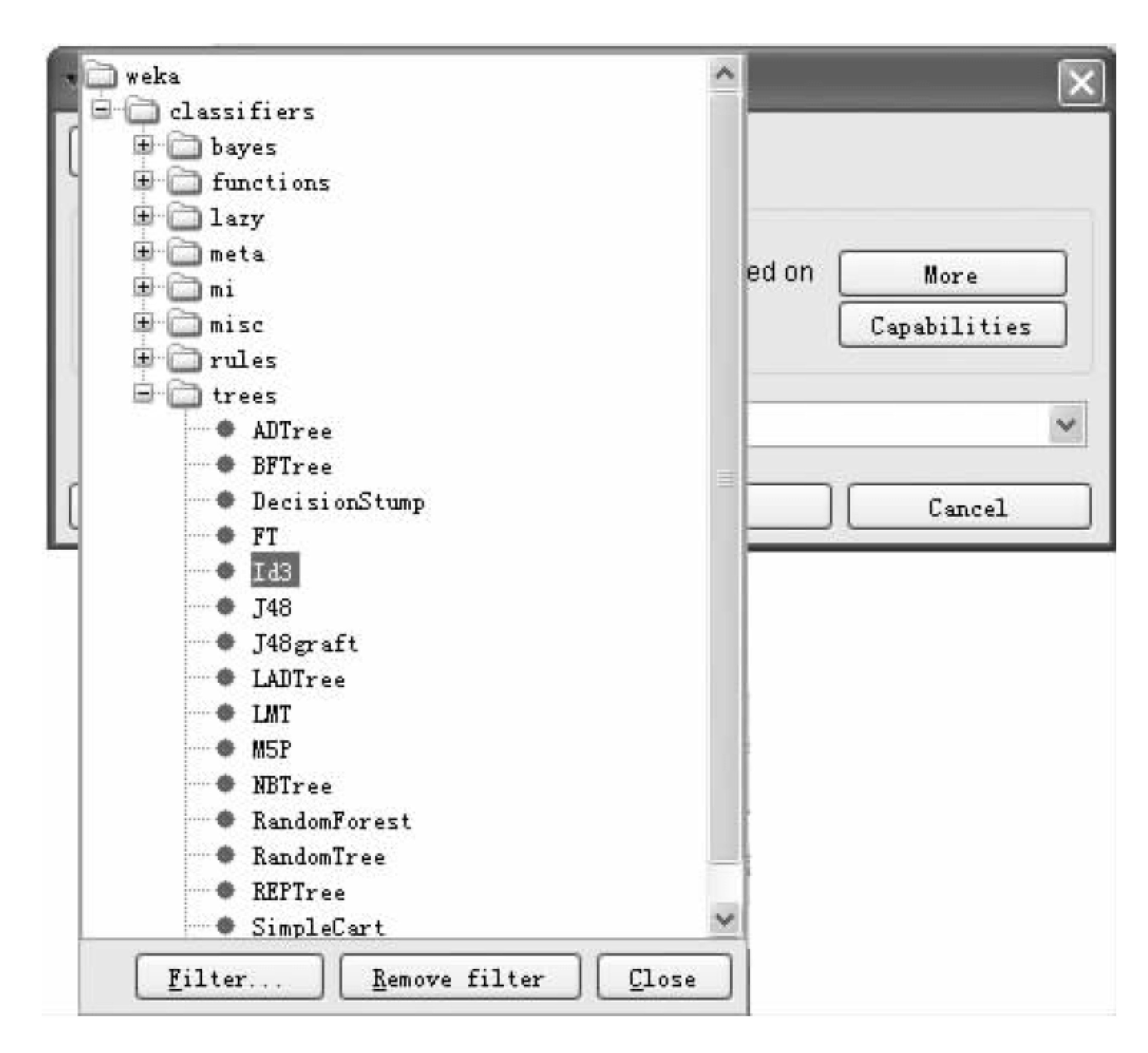


图 5-5　选择算法

（5）选择 ID3 算法和 BayesNet 算法，用这两种算法对 HumanResource 数据集进行分类，并对比分类结果。设置完成后的页面状态如图 5-6 所示。

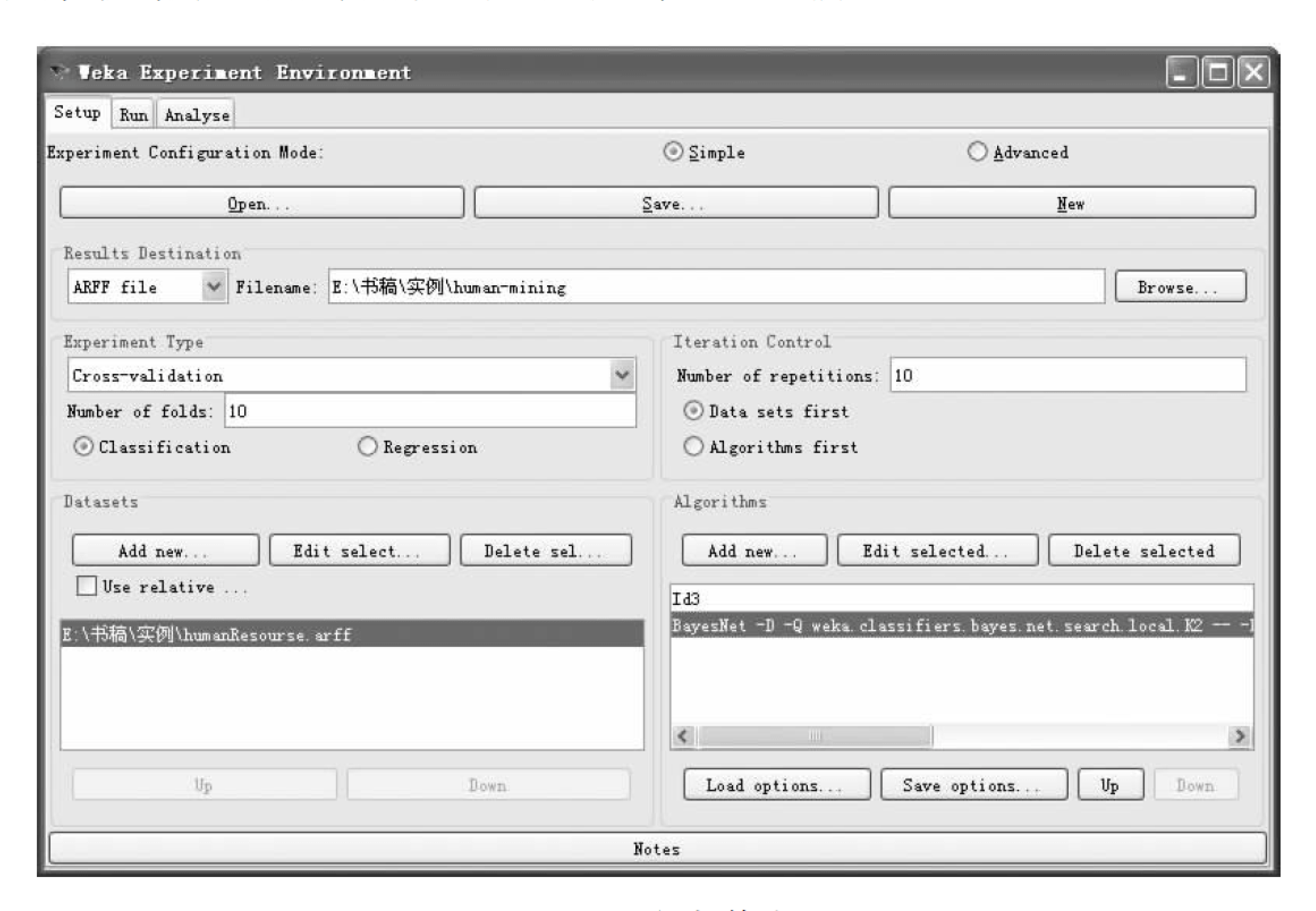


图 5-6　添加算法

（6）从任务面板选择区域单击 Run 标签，打开运行面板，并在该面板内单击 Start 按钮，Weka 开始用上面选择的两个算法对数据集进行处理，并同时在 Log 区域显示处理的开始结束时间，以及错误数，如图 5-7 所示。如果显示处理完成，并且没有错误发生（There were 0 errors），则表明处理已经正确完成。

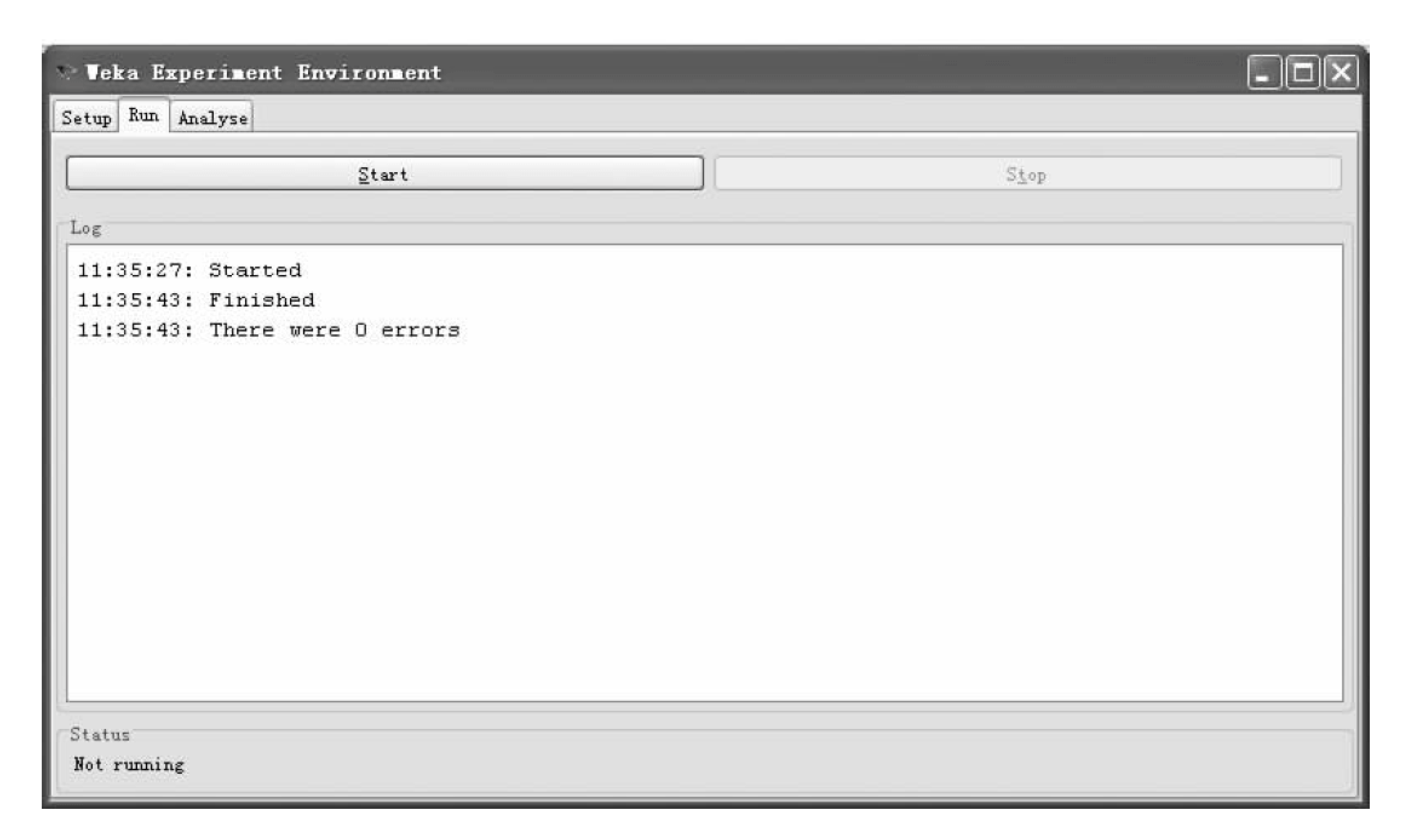


图 5-7　运行算法

（7）单击 Analyse 标签转到结果分析面板，并在该面板右上方单击 Experiment 按钮，表示结果分析的来源为刚刚运行的实验结果。也可以在任何时候单击 File 按钮，从实验配置中设置的结果文件将实验结果装入，如图 5-8 所示。

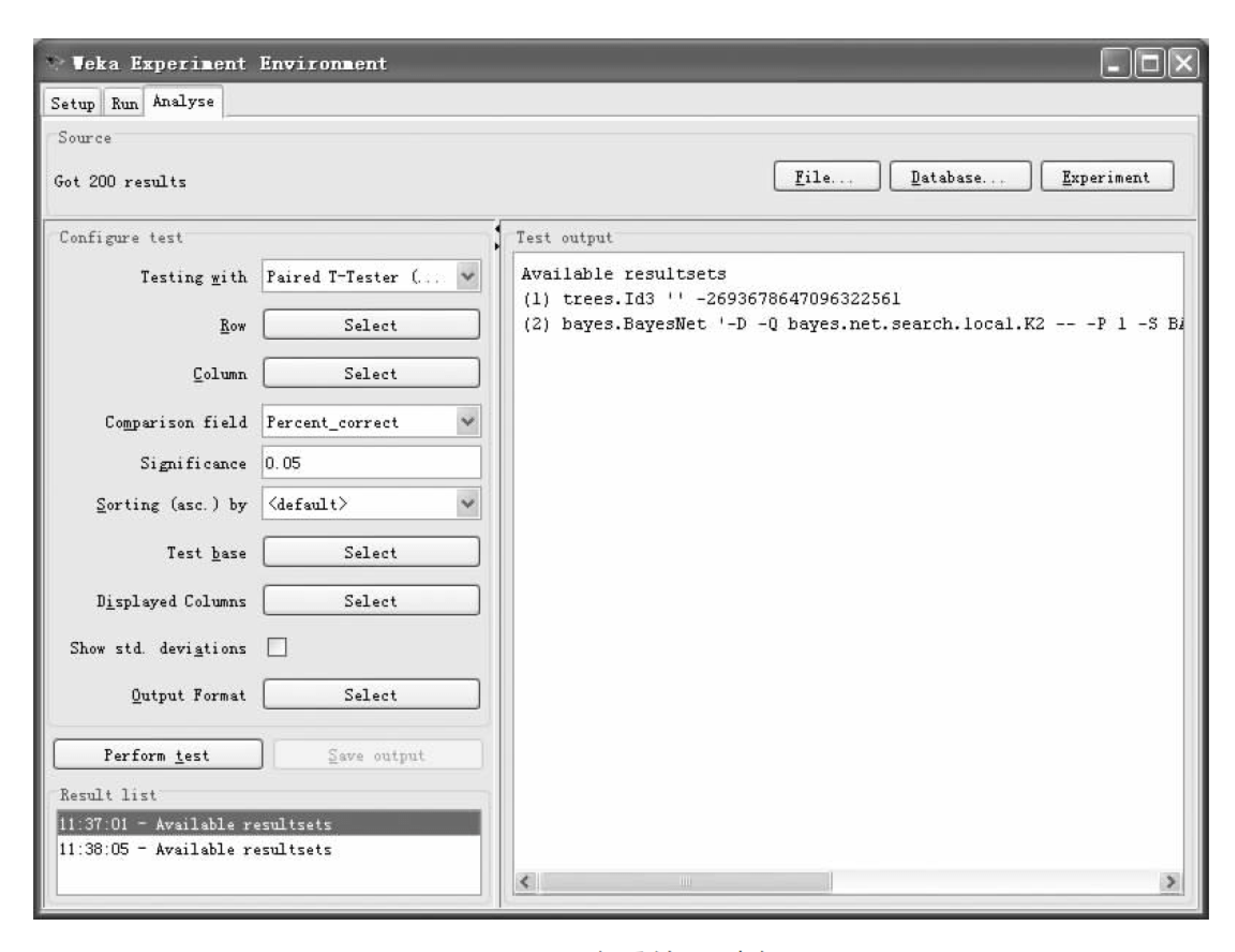


图 5-8　查看结果列表

（8）图 5-8 中的 Result list 列表框中有两个实验结果，第一个是通过 Experiment 按钮装入的；第二个是通过 File 按钮装入的，选中其中一个，就会在 Test output 区域显示该实验的信息。

可以通过左边的 Configure test 区域选择各种需要比较的参数。

单击 Perform test 按钮，将在 Test output 区域显示测试结果，如图 5-9 所示。

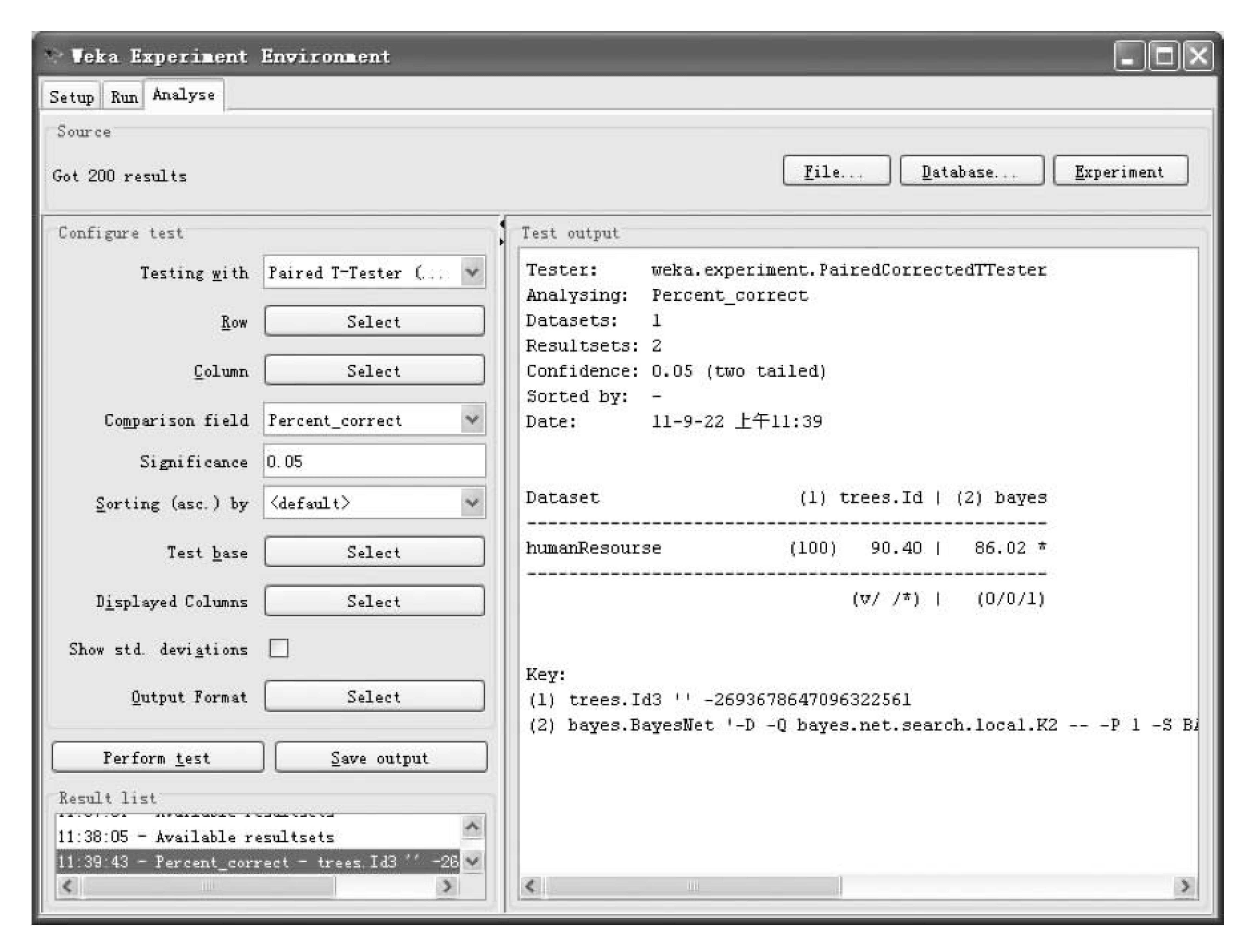


图 5-9　显示测试结果

结果显示了测试模型的基本情况，以及测试结果，包括使用的数据集名称以及两种算法分类结果比较。

Configure test 区域中提供了一些选项，用户可以通过这些选项，选择感兴趣的比较条件。例如，本例主要关心使用两类算法分类的准确性比较，可以在 Comparison field 下拉列表框中选择比较条件，如正确分类的百分比、分类不正确的百分比、没有分类数据的百分比、分类误差等等。

对于本例中的数据，ID3 算法在各方面都优于 bayes 算法，所以选择 ID3 算法作为本问题的分类器。

5.4　案例小结

Experimenter 模块的作用是比较多个候选模型的优劣，帮助用户选择挖掘模型。本实例首先介绍了对来源比较复杂的数据源进行预处理所需要的几个关键步骤，然后在 Weka 的 Experimenter 模块中建立一个实验，通过运行该实验对选择的两个分类模型(ID3 模型和 bayes 模型)进行比较，最终确定一个最优的挖掘模型。

读者可以从本章获得的知识有两个方面，一是了解人力资源数据挖掘的特点，数据预处理方法和挖掘目的；二是学习 Weka 的 Experimenter 模块的使用方法。

实例 6　基于贝叶斯方法的证券客户流失预警分析

6.1　任务描述

随着证券市场竞争日益加剧，证券市场已经由佣金战以及价格战转向服务战，而服务战的核心便是以客户为中心展开的。由于客户保持的运营成本远小于客户的新增开发成本，所以如何预防客户流失被越来越多的券商所关注。在其他的行业如移动通信行业，大量的实践经验已经证明：客户保持工作的最佳时机是在其未流失时，其原因在于已流失客户的回流阻力大，策反工作成本高且效果差。在券商经纪业务中也同样存在着类似的问题。券商们在面对其庞大的客户群时，不可能对每个客户都去做特殊的保护工作。这就需要券商建立相应的流失预警机制，通过对将要流失的高价值客户进行预测，及时了解他们的需求，投入一定的资源，并通过针对性的挽留工作避免其流失。这样可以提高效率，而且可以大大节省成本，获得可观的收益。

根据证券从业人员经验，以下 9 个属性是可能造成客户流失的重要因素：

- 客户级别(khjb)；
- 资金转出率(zjzcl)；
- 账户空置时间(zhkzsj)；
- 开户时间(khsj)；
- 客户佣金率(khyjl)；
- 客户月资产收益率(khyzcsyl)；
- 2010 年佣金贡献(nyjgx)；
- 营业部竞争压力(yybjzyl)；
- 地域(dy)。

某券商根据以上影响因素，收集了 2000 名客户的相关数据，其中 1000 名为销户客户，1000 名为未销户客户，实际部分数据如表 6-1 所示。表中，xiaohu＝b 表示未销户客户；xiaohu＝c 表示销户客户。

表 6-1　证券客户数据

1	yybjzyl	khsj	khjb	zjzcl	zhkzsj	khyjl	nyjgx	khyzcsyl	dy	xiaohu
2	b	f	f	a	a	a	a	a	b	b
3	b	f	f	e	a	f	b	f	b	b
4	b	f	f	d	d	f	c	f	b	b
5	b	f	f	a	a	a	a	a	b	b
6	b	f	e	e	d	c	c	e	b	b
7	b	f	f	a	a	a	a	a	b	b
8	b	f	f	d	a	f	b	e	b	b
9	b	f	f	a	a	a	a	a	b	b
10	b	f	f	a	a	a	a	a	b	b
11	b	e	f	a	a	a	a	a	b	b
12	b	f	f	a	a	a	a	a	b	b
13	b	f	f	a	a	a	a	a	b	b
14	b	f	f	a	a	a	a	a	b	b
15	b	e	f	a	a	a	a	a	b	b

（1）试根据这些数据建立证券客户流失预警模型。

（2）评估预警模型的可行性。

6.2 技术原理

朴素贝叶斯分类器是贝叶斯分类模型中一种最简单、有效而且在实际使用中很成功的分类器，朴素贝叶斯分类基于贝叶斯定理，在实际运用中降低了贝叶斯网络构建的复杂性。分类算法的比较研究发现，朴素贝叶斯分类算法可以与判定树和神经网络分类算法相媲美；用于大型数据库分析，朴素贝叶斯分类也已表现出高准确率与高速度，而且已经成功地应用于聚类、分类等数据挖掘任务中。

6.2.1 朴素贝叶斯分类算法

（1）每个数据样本用一个 n 维特征向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 表示，属性 $A_1,A_2,\cdots,A_n$ 描述对样本的 n 个度量。

（2）假定有 m 个类 $C_1,C_2,\cdots,C_m$。给定一个未知的数据样本 X（即没有类标号），分类法将预测 X 属于具有最高后验概率（条件 X 下）的类。即朴素贝叶斯分类将未知的样本分配给类 C_i，当且仅当：

$$p\left(\frac{C_i}{X}\right)>p\left(\frac{C_j}{X}\right),\quad 0\leqslant j\leqslant m,j\neq i$$

于是可以最大化 $p\left(\frac{C_i}{X}\right)$，其中 $p\left(\frac{C_i}{X}\right)=\dfrac{p\left(\frac{X}{C_i}\right)p(C_i)}{p(X)}$。

（3）由于 $p(X)$ 对于所有类为常数，只需要 $p\left(\frac{X}{C_i}\right)p(C_i)$ 最大即可。

若类的先验概率未知，则通常假定这些类是等概率的，即 $p(C_1)=p(C_2)=\cdots=p(C_m)$。据此只需对 $p\left(\frac{X}{C_i}\right)$ 最大化。

若类的先验概率已知，则最大化 $p\left(\frac{X}{C_i}\right)p(C_i)$。类的先验概率可以用 $p(C_i)=\frac{s_i}{s}$ 计算。其中，s_i 是类 C_i 中的训练样本数；s 是训练样本总数。

（4）给定具有许多属性的数据集，计算 $p\left(\frac{X}{C_i}\right)$ 的开销可能非常大。为降低计算 $p\left(\frac{X}{C_i}\right)$ 的开销，可以做类条件独立的朴素假定，即给定样本的类标号，假定属性值条件地相互独立，即属性间不存在依赖关系。这样，

$$p\left(\frac{X}{C_i}\right)=\prod_{k=1}^{n}p\left(\frac{x_k}{C_i}\right)\tag{6-1}$$

概率 $P\left(\frac{x_1}{C_i}\right),P\left(\frac{x_2}{C_i}\right),\cdots,P\left(\frac{x_n}{C_i}\right)$ 可以由训练样本估值，其中：

- 如果 A_k 是离散型属性，则 $P\left(\frac{X_k}{C_i}\right)=\frac{s_{ik}}{s_i}$；$s_{ik}$ 是在属性 A_k 上具有值 x_k 的类 C_i 的训练

样本数，而 s_i 是 C_i 中的训练样本数。

- 如果 A_k 是连续型属性，则通常假定该属性服从高斯分布。因而

$$p\left(\frac{x_k}{C_i}\right)=g(x_k,\mu_{C_i},\sigma_{C_i})=\frac{1}{\sqrt{2\pi}\sigma_{C_i}}e^{-\frac{(x-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \tag{6-2}$$

其中，给定类 C_i 的训练样本属性 A_k 的值，$g(x_k,\mu_{C_i},\sigma_{C_i})$ 是属性 A_k 的高斯密度函数，而 μ_{C_i}，σ_{C_i} 分别为平均值和标准差。

(5) 对每个类 C_i，计算 $P\left(\frac{X}{C_i}\right)P(C_i)$。样本 X 被指派到类 C_i，当且仅当：

$$P\left(\frac{X}{C_i}\right)P(C_i)>P\left(\frac{X}{C_j}\right)P(C_j),\quad 1\leqslant j\leqslant m,j\neq i$$

换言之，X 被指派到使 $P\left(\frac{X}{C_i}\right)P(C_i)$最大的类 C_i。

6.2.2 朴素贝叶斯分类举例

给定训练数据如表 6-2 所示，数据样本用属性 age、income、student 和 credit_rating 描述。类标号属性 buys_computer 具有两个不同值(即{yes,no})。给定一个没有类标号的数据样本 X=(age="<=30", income="medium", student="yes", credit_rating="fair")，下面使用朴素贝叶斯分类预测这个数据样本的类标号。

表 6-2 AllElectronics 顾客数据库训练数据元组

RID	age	income	student	Credit_rating	Class: buys_computer
1	<=30	high	no	fair	No
2	<=30	high	no	excellent	No
3	31…40	high	no	fair	Yes
4	>40	medium	no	fair	Yes
5	>40	low	yes	fair	Yes
6	>40	low	yes	excellent	No
7	31…40	low	yes	excellent	Yes
8	<=30	medium	no	fair	No
9	<=30	low	yes	fair	Yes
10	>40	medium	yes	fair	Yes
11	<=30	medium	yes	excellent	Yes
12	31…40	medium	no	excellent	Yes
13	31…40	high	yes	fair	Yes
14	>40	medium	no	excellent	No

设 C_1 对应于类 buys_computer="yes"，而 C_2 对应于类 buys_computer="no"。根据前面的讲述，需要最大化 $P\left(\frac{X}{C_i}\right)P(C_i)$，$i=1,2$。

每个类的先验概率 $P(C_i)$ 可以根据训练样本计算：

P(buys_computer="yes")=9/14=0.643

P(buys_computer="no")=5/14=0.357

为计算 $p(X/C_i)$,$i=1,2$,计算下面的条件概率:

P(age="<30"|buys_computer="yes")=2/9=0.222

P(age="<30"|buys_computer="no")=3/5=0.600

P(income="medium"|buys_computer="yes")=4/9=0.444

P(income="medium"|buys_computer="no")=2/5=0.400

P(student="yes"|buys_computer="yes")=6/9=0.667

P(student="yes"|buys_computer="no")=1/5=0.200

P(credit_rating="fair"|buys_computer="yes")=6/9=0.667

P(credit_rating="fair"|buys_computer="no")=2/5=0.400

使用以上概率,得到:

P(X|buys_computer="yes")=0.222×0.444×0.667×0.667=0.044

P(X|buys_computer="no")=0.600×0.400×0.200×0.400=0.019

P(X|buys_computer="yes") P(buys_computer="yes")=0.044×0.643=0.028

P(X|buys_computer="no") P(buys_computer="no")=0.019×0.357=0.007

因此,对于样本 X,由于 0.028>0.007,朴素贝叶斯分类预测 buys_computer="yes"。

6.3 具体实现

(1) 选择"开始"→"所有程序"→Weka 3.6.5→Weka 3.6 命令,如图 6-1 所示。



图 6-1 打开 Weka 软件

(2) 在打开的文件中,单击 Explorer 按钮,如图 6-2 所示。



图 6-2 打开 Explorer 应用

（3）单击 Open file 按钮，选择要打开的文件 zhqkehuliushi.csv，并单击“打开”按钮，如图 6-3 所示。

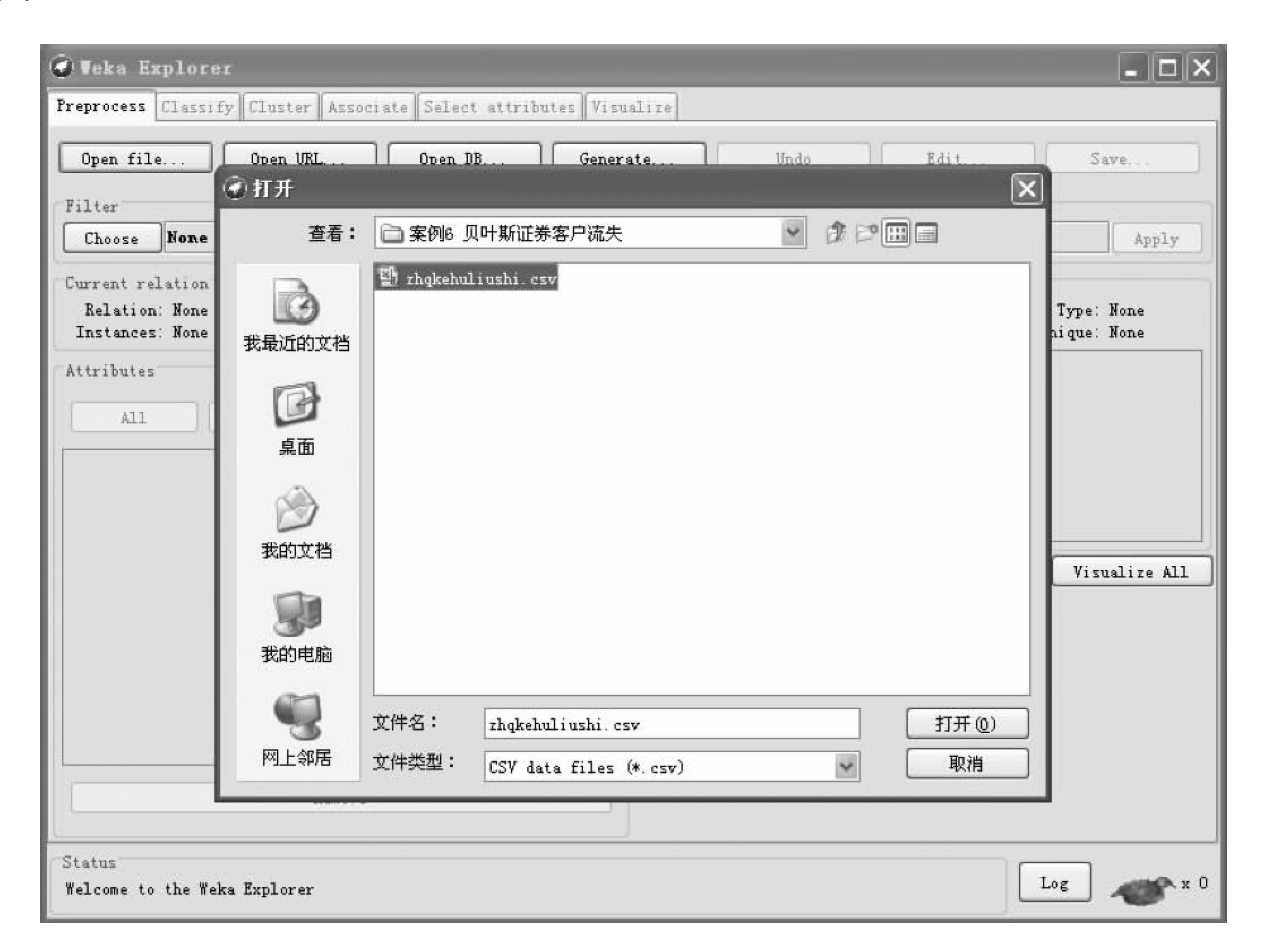


图 6-3　打开数据文件

（4）在打开如图 6-4 所示的界面中，知道 zhqkehuliushi 数据集中共有 2000 个实例，每个实例有 10 个属性。选中某个属性，可以查看 2000 个实例关于这个属性值的最小值、最大值、均值和标准差等信息。然后单击 Classify 标签，并单击 Choose 按钮，如图 6-4 所示。

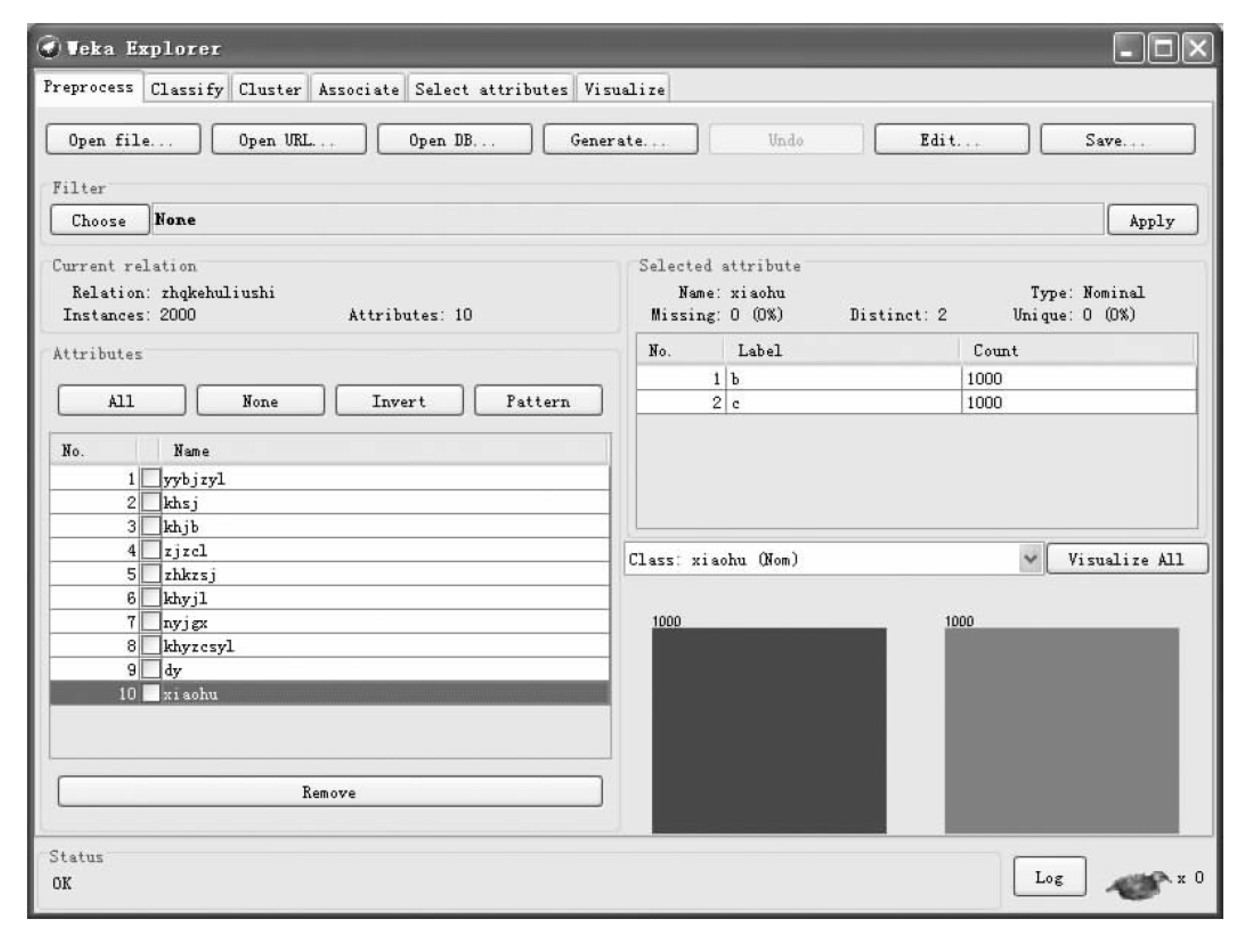


图 6-4　查看数据特征

(5) 如图 6-5 所示，选择 NaiveBayes 分类器，单击 Close 按钮。

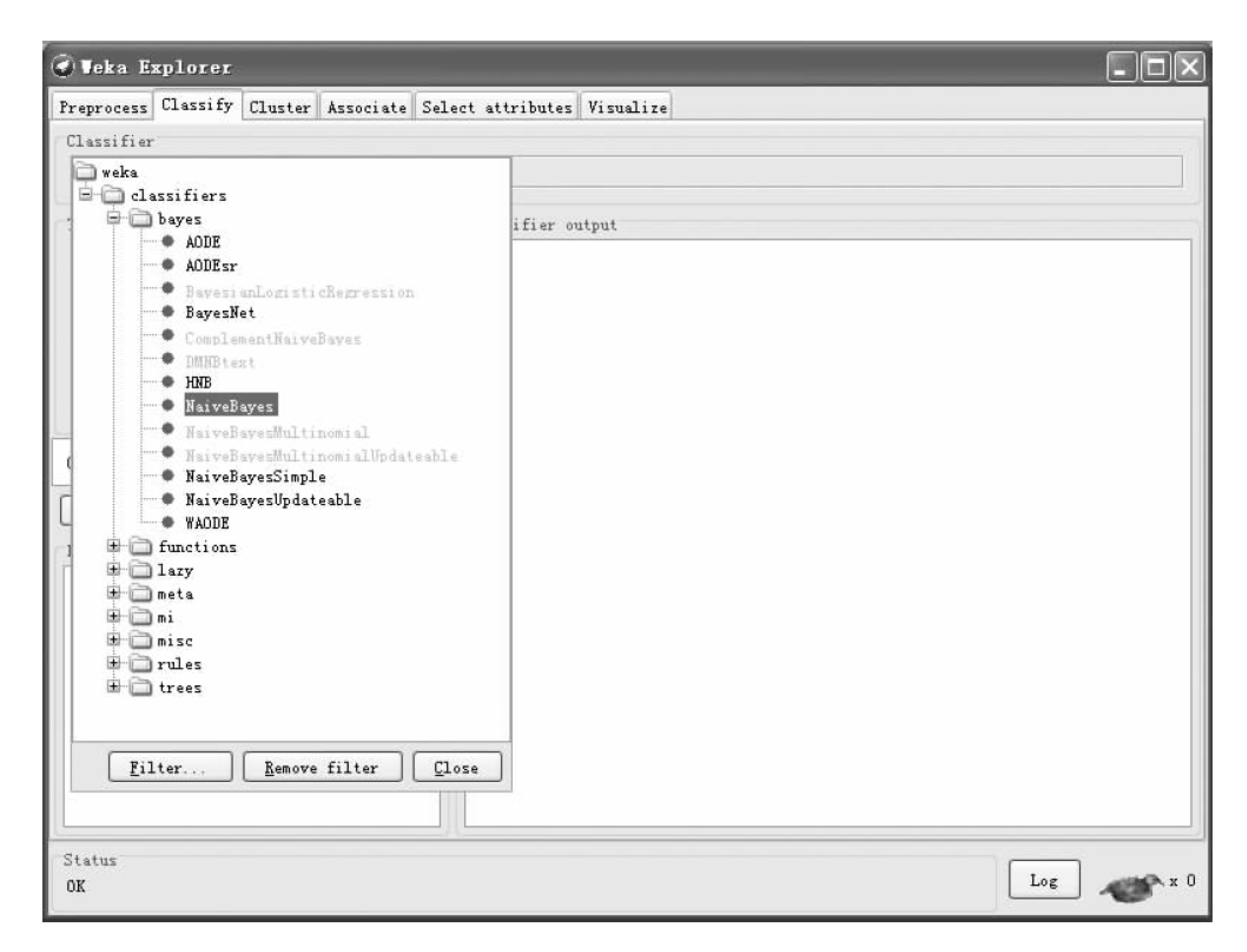


图 6-5 选择分类方法

(6) 双击 NaiveBayes 可以对算法的参数进行设置，这里选择默认参数，如图 6-6 所示。

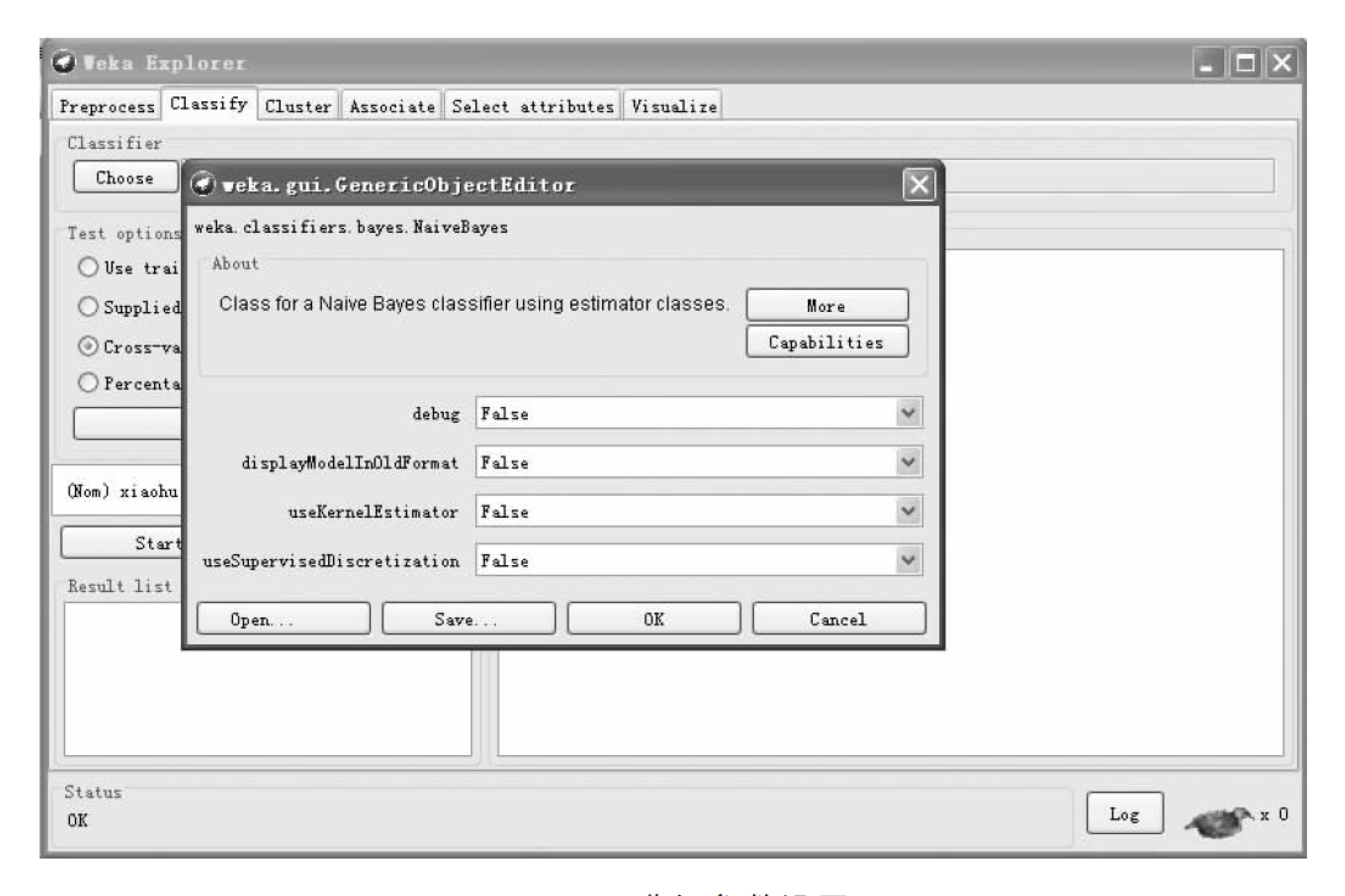


图 6-6 进行参数设置

(7) 测试方法选择 10 折交叉验证，单击 Start 按钮，Weka 软件显示运行结果，可知正确率为 86.75%，如图 6-7 所示。

(8) 在结果显示中还可以看到，1000 名未销户客户有 855 名预测为未销户客户，145 名预测为销户客户；1000 名销户客户有 880 名预测为销户客户，120 名预测为未销户客户，如图 6-8 所示。

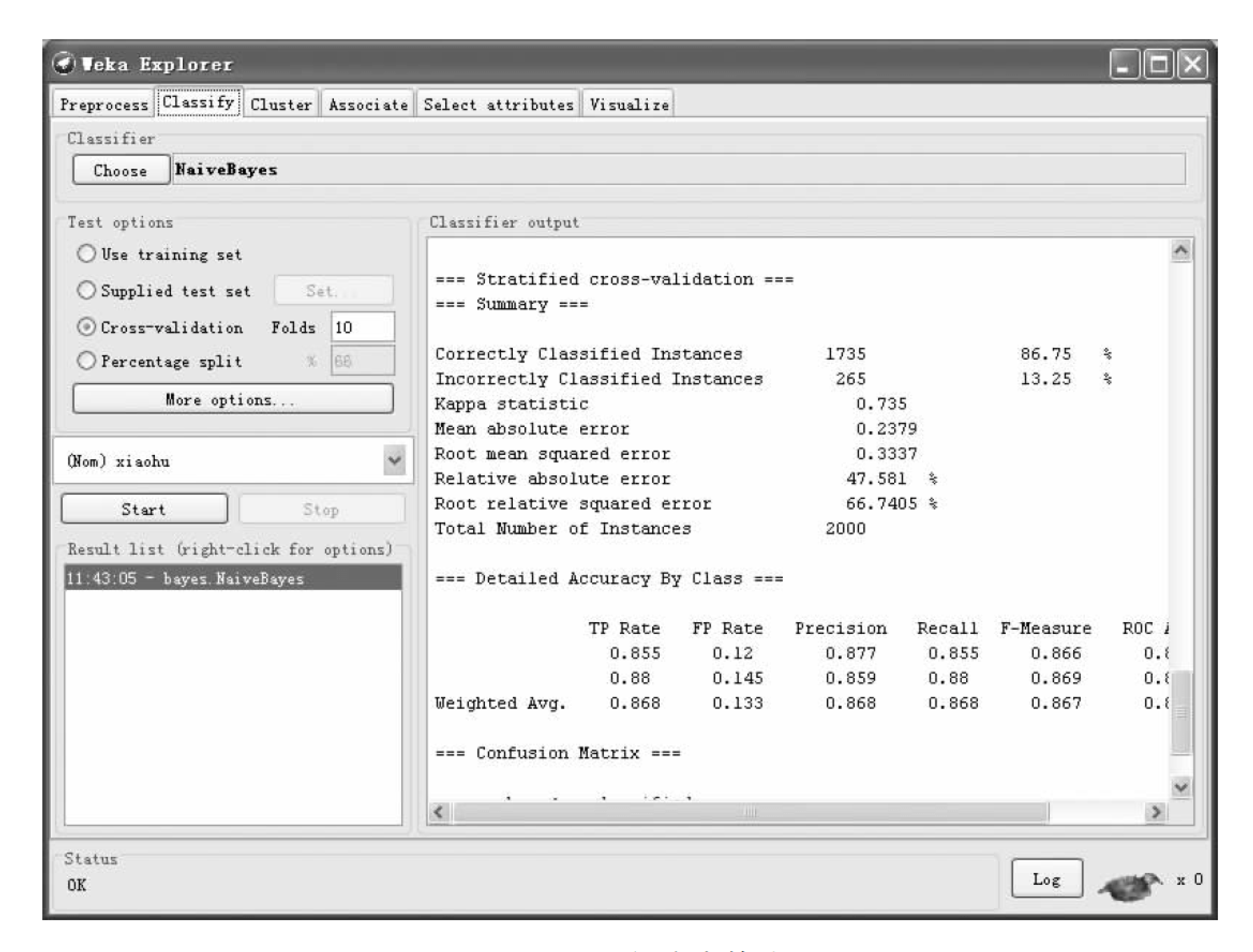


图 6-7　运行分类算法

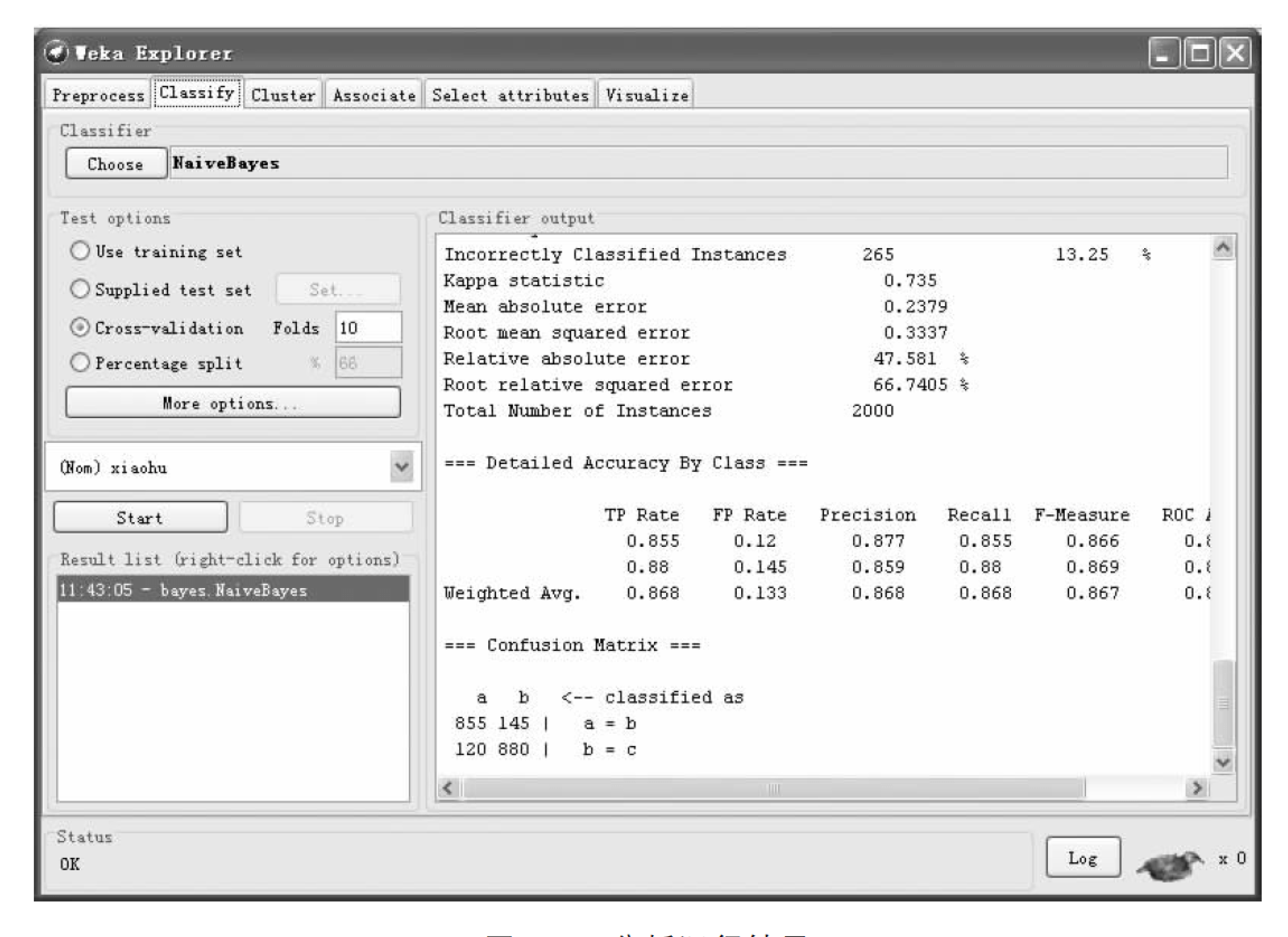


图 6-8　分析运行结果

（9）如图 6-9 所示，可知正确率为 86.75％，利用朴素贝叶斯模型已基本能达到券商的要求。

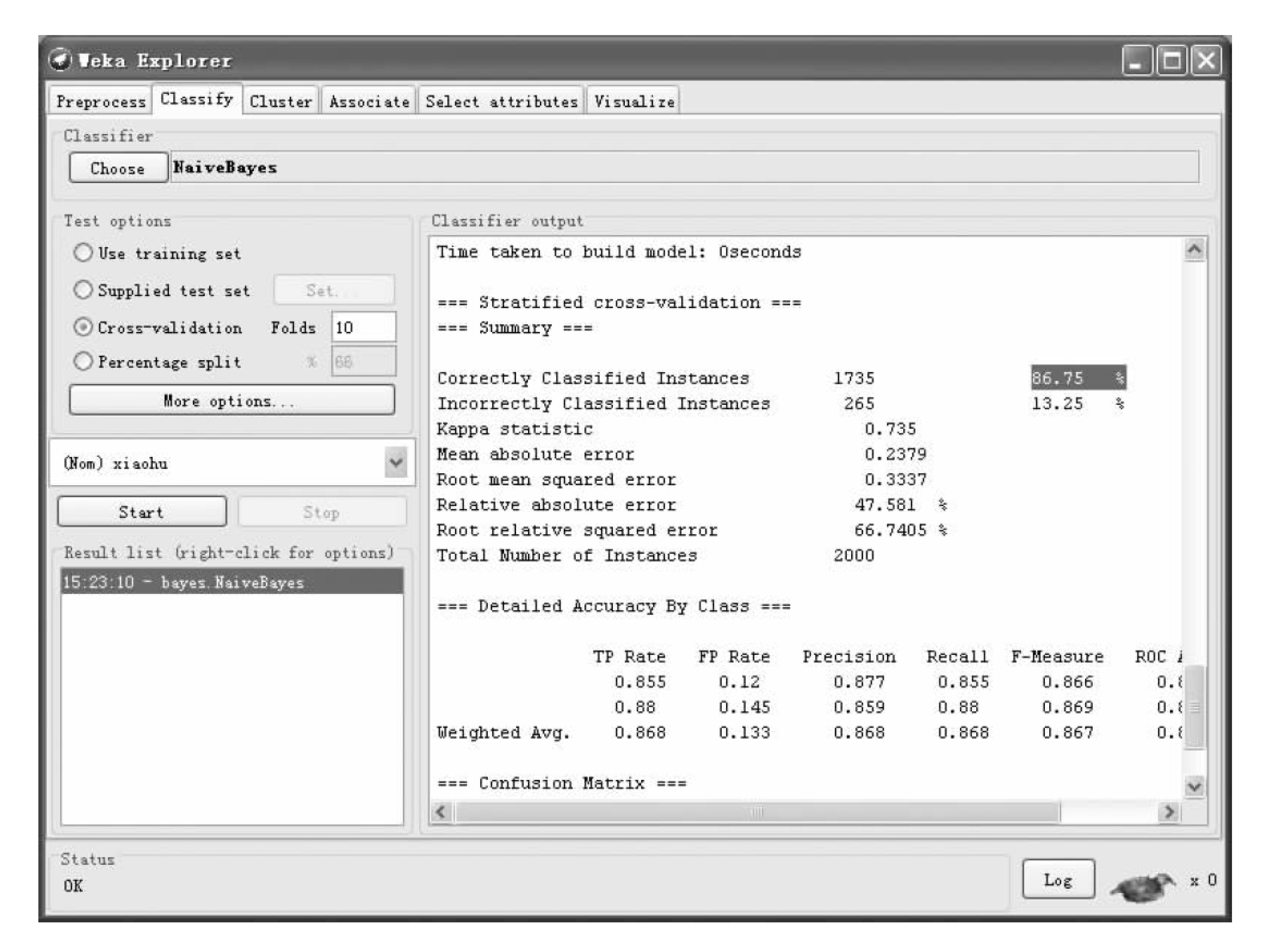


图 6-9 查看正确率

6.4 案例小结

朴素贝叶斯分类基于贝叶斯定理,已经成功地应用于聚类、分类等数据挖掘任务中。分类算法的比较研究发现,朴素贝叶斯分类算法可以与判定树和神经网络分类算法相媲美;用于大型数据库分析,朴素贝叶斯分类也已表现出高准确率与高速度。本案例结合证券客户流失实际数据,利用 Weka 软件中提供的朴素贝叶斯分类方法,建立了证券客户流失预警模型,取得了较高的客户流失预警正确率。本案例可以在补充测试集的情况下,进一步检验预警模型的可行性。

实例 7　基于人工神经网络方法的信贷数据分析

7.1　任 务 描 述

本案例为“某商业银行信贷”数据库，其中记录了客户的背景数据以及贷款情况，包括“客户基本情况表”、“分行代码表”和“贷款余额表”三张基本表。其中，客户基本情况包括“客户代码”、“客户名称”、“客户类型”、“经济性质”、“隶属关系”、“法人资格”、“客户状态”和“重点标志”共 8 个属性。其中，“客户名称”用代号区分各个单位，没有具体提及真实单位名称。这些属性都是银行发现客户背景与不良贷款记录之间关系的主要依据。“贷款余额表”记录了客户的贷款及归还情况。

不良贷款可以界定为银行投放贷款后形成的信贷资产中不符合安全性、流动性、盈利性原则，处于逾期、呆滞或呆账状态，而使银行资产风险加大并面临资本损失的那部分贷款。按照人民银行的规定，不良贷款可以分成 5 类：正常、关注、次级、可疑和损失。本案例的任务就是发现具有哪些背景的用户更容易产生不良贷款，从而对贷款去向进行监督，避免损失，也可以帮助银行为优质客户提供更好的服务。

7.2　技 术 原 理

以数学和物理方法以及从信息处理的角度对人脑神经网络进行抽象，并建立某种简化模型，称为人工神经网络(Artificial Neural Network，ANN)。在模式识别、系统辨识、信号处理、自动控制、组合优化、预测预估、故障诊断、数据挖掘、医学和经济学等领域，人工神经网络已经成功解决了许多现代计算机难以解决的实际问题，表现出良好的智能特性和潜在的应用前景。

人工神经网络的特点和优势主要表现在以下 3 个方面：第一，具有自学习功能。例如，实现图像识别时，只要先把不同的图像样本和对应的识别结果输入人工神经网络，网络就会通过自学习功能，慢慢学习识别类似的图像。第二，具有联想存储功能。人工神经网络的反馈网络可以实现这种联想。例如，经过训练的神经网络可以从“眼睛”特征恢复整个人脸图像，这叫做自联想，从“勺子”联系出“筷子”、“碗”等，这叫做互联想。第三，具有高速寻找优化解的能力。寻找某个复杂问题的优化解往往需要很大的计算量，利用针对特定问题而设计的反馈型人工神经网络，发挥计算机的高速运算能力，可以很快找到优化解。

BP(Back Propagation)神经网络是迄今为止应用最为广泛的神经网络，现将该网络的一些基本知识点进行简单回顾，以便读者理解该案例的应用。

7.2.1　BP 神经网络结构

BP 神经网络不仅有输入结点、输出结点，而且还有一层或多层隐含结点，神经元的变换函数采用(0,1)S 型函数，如图 7-1 所示。

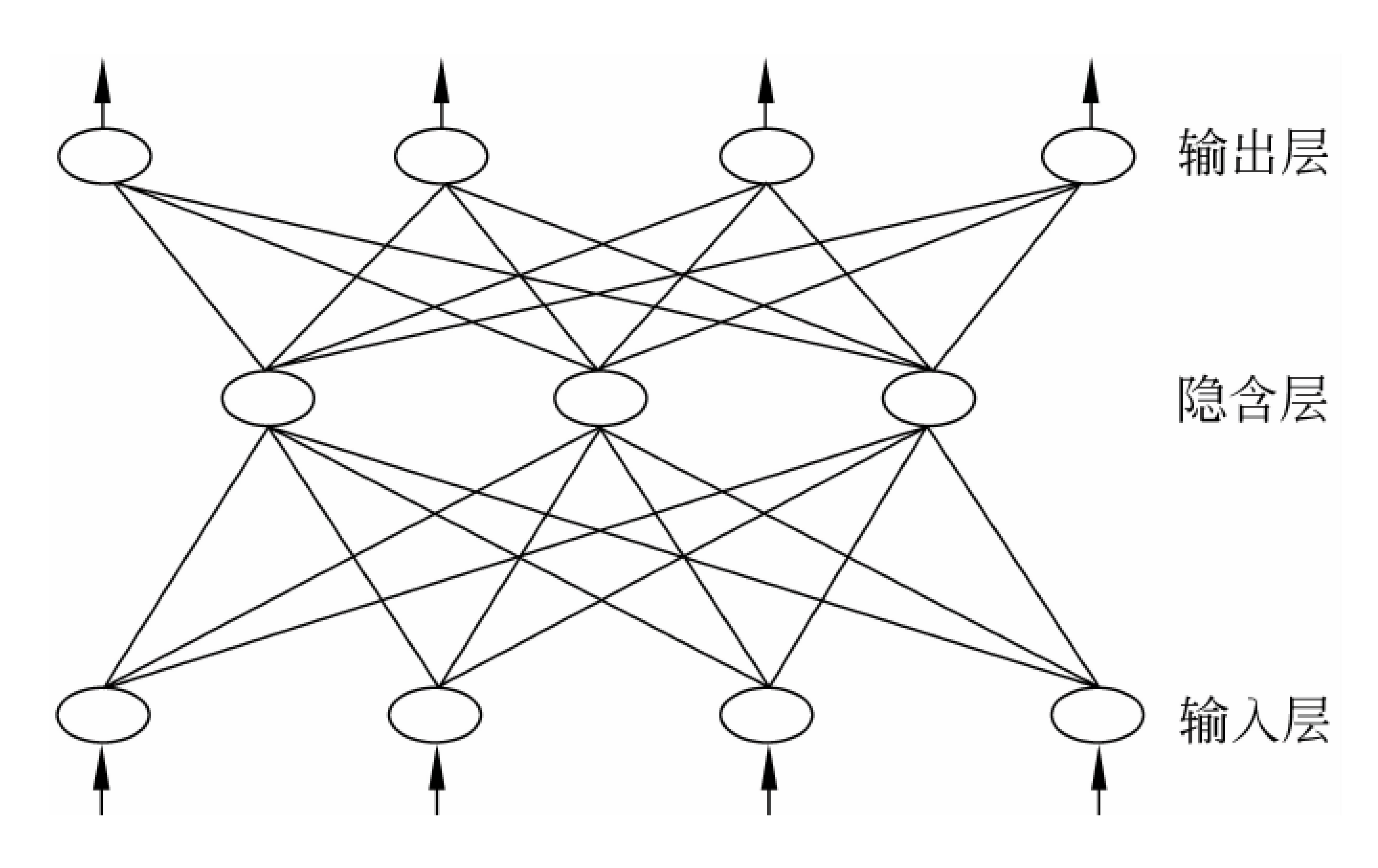


图 7-1　BP 神经网络示意图

在图 7-1 中，最下面的层为输入层，第 Q 层为输出层，中间各层为隐含层，设第 q 层（$q=1,2,\cdots,Q$）的神经元个数为 n_q，输入到第 q 层的第 i 个神经元的连接权系数为 $\tilde{\omega}_{ij}^{(q)}$（$i=1,2,\cdots,n_q$；$j=1,2,\cdots,n_{q-1}$）。该网络的输入输出变换关系为

$$s_i^{(q)} = \sum_{j=0}^{n_q-1} \tilde{\omega}_{ij}^{(q)} x_j^{(q-1)} \quad (x_0^{(q-1)} = \theta_i^{(q)}, \tilde{\omega}_{i0}^{(q)} = -1)$$

$$x_i^{(q)} = f(s_i^{(q)}) = \frac{1}{1+\mathrm{e}^{-\mu s_i^{(q)}}}$$

$$i = 1,2,\cdots,n_q, \quad j = 1,2,\cdots,n_{q-1}, \quad q = 1,2,\cdots,Q \tag{7-1}$$

7.2.2　BP 神经网络学习算法

在 BP 神经网络中，输入信号是从输入层到隐层再到输出层传递的。最后一个隐层与输出层之间的连接权是输出误差的显函数，而其他层之间的连接权则是输出误差的隐函数。如果神经元的作用函数是连续可微的，那么每一连接权对输出误差的影响都可以由误差对权值的偏导数定量的描述。此时，如果把权值按照梯度的反方向修正则可以使误差减小。这种思想便是误差反向传播（BP 算法）的本质。详细计算方法如下：

设取拟合误差的代价函数为

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{n_Q}(d_{pi} - x_{pi}^{(Q)})^2 = \sum_{p=1}^{P} E_p \tag{7-2}$$

即

$$E_p = \frac{1}{2}\sum_{i=1}^{n_Q}(d_{pi} - x_{pi}^{(Q)})^2 \tag{7-3}$$

问题是如何调整连接权系数以使代价函数最小。优化计算的方法很多，比较典型的是采用一阶梯度法，即最速下降法。该方法的关键是计算优化目标函数（即上述的误差代价函数）E 对寻优参数的一阶导数。依次从输出层开始计算如下：

$$\frac{\partial E}{\partial \tilde{\omega}_{ij}^{(q)}} \quad (q = Q, Q-1, \cdots, 1)$$

由于

$$\frac{\partial E}{\partial \tilde{\omega}_{ij}^{(q)}} = \sum_{p=1}^{P} \frac{\partial E_p}{\partial \tilde{\omega}_{ij}^{(q)}}$$

所以应着重讨论 $\dfrac{\partial E_p}{\partial \tilde{\omega}_{ij}^{(q)}}$ 的计算。

对于第 Q 层有

$$\frac{\partial E_p}{\partial \tilde{\omega}_{ij}^{(Q)}} = \frac{\partial E_p}{\partial x_{pi}^{(Q)}} \frac{\partial x_{ij}^{(Q)}}{\partial s_{pi}^{(Q)}} \frac{\partial s_{ipi}^{(Q)}}{\partial \tilde{\omega}_{ij}^{(Q)}} = -(d_{pi} - x_{pi}^{(Q)}) f'(s_{pi}^{(Q)}) x_{pi}^{(Q-1)} = -\delta_{pi}^{(Q)} x_{pi}^{(Q-1)} \tag{7-4}$$

其中

$$\delta_{pi}^{(Q)} = -\frac{\partial E_p}{\partial s_{pi}^{(Q)}} = (d_{pi} - x_{pi}^{(Q)}) f'(s_{pi}^{(Q)})$$

$x_{pi}^{(Q)}$、$s_{pi}^{(Q)}$ 及 $x_{pi}^{(Q-1)}$ 表示利用第 p 组输入样本所算得的结果。

对于第 $Q-1$ 层有

$$\begin{aligned}\frac{\partial E_p}{\partial \tilde{\omega}_{ij}^{(Q-1)}} &= \frac{\partial E_p}{\partial x_{pi}^{(Q-1)}} \frac{\partial x_{ipi}^{(Q-1)}}{\partial \tilde{\omega}_{ij}^{(Q-1)}} = \left(\sum_{k=1}^{n_Q} \frac{\partial E_p}{\partial s_{pk}^{(Q)}} \frac{\partial s_{pk}^{(Q)}}{\partial x_{pi}^{(Q-1)}}\right) \frac{\partial x_{pi}^{(Q-1)}}{\partial s_{pi}^{(Q-1)}} \frac{\partial s_{pi}^{(Q-1)}}{\partial \tilde{\omega}_{ij}^{(Q-1)}} \\ &= \left(\sum_{k=1}^{n_Q} \delta_{pk}^{(Q)} \tilde{\omega}_{ki}^{(Q)}\right) f'(s_{ipi}^{(Q-1)}) x_{pj}^{(Q-2)} = -\delta_{pi}^{(Q-1)} x_{pj}^{(Q-2)}\end{aligned} \tag{7-5}$$

其中

$$\delta_{pi}^{(Q-1)} = -\frac{\partial E_p}{\partial s_{pi}^{(Q-1)}} = \left(\sum_{k=1}^{n_Q} \delta_{pk}^{(Q)} \tilde{\omega}_{ki}^{(Q)}\right) f'(s_{ipi}^{(Q-1)})$$

显然,它是反向递推计算的公式,即首先计算出 $\delta_{pk}^{(Q)}$ 然后再由上式递推计算出 $\delta_{pi}^{(Q-1)}$。依次类推,可继续反向递推计算出 $\delta_{pi}^{(q)}$ 和 $\dfrac{\partial E_p}{\partial \tilde{\omega}_{ij}^{(q)}}$($q=Q-2,\cdots,1$)。从上式看出,在 $\delta_{pi}^{(q)}$ 的表达式中包含了导数项 $f'(s_{pi}^{(q)})$,由于 BP 网络使用 S 形函数,所以其导数可求得如下:

$$x_{pi}^{(q)} = f(s_{pi}^{(q)}) = \frac{1}{1+\mathrm{e}^{-\mu s_{pi}^{(q)}}}$$

$$f'(s_{pi}^{(q)}) = \frac{\mu \mathrm{e}^{-\mu s_{pi}^{(q)}}}{(1+\mathrm{e}^{-\mu s_{pi}^{(q)}})^2} = \mu f(s_{pi}^{(q)})[1-f(s_{pi}^{(q)})] = \mu x_{pi}^{(q)}(1-x_{pi}^{(q)}) \tag{7-6}$$

最后可归纳出 BP 网络的学习算法如下:

$$\begin{aligned}&W_{ij}^{(q)}(k+1) = \tilde{\omega}_{ij}^{(q)}(k) + \alpha D_{ij}^{(q)}(k+1), \quad \alpha > 0 \\ &D_{ij}^{(q)} = \sum_{p=1}^{P} \delta_{pi}^{(q)} x_{pj}^{(q-1)} \\ &\delta_{pi}^{(q)} = \left(\sum_{k=1}^{n_q+1} \delta_{pk}^{(q+1)} \tilde{\omega}_{ki}^{(q+1)}\right) \mu x_{pi}^{(q)} (1 - x_{pi}^{(q)}) \\ &\delta_{pi}^{(Q)} = (d_{pi} - x_{pi}^{(Q)}) \mu x_{pi}^{(Q)} (1 - x_{pi}^{(q)}) \\ &q = Q, Q-1, \cdots, 1; \quad i = 1, 2, \cdots, n_q; \quad j = 1, 2, \cdots, n_{q-1}\end{aligned} \tag{7-7}$$

BP 网络由于其很好的逼近非线性映射的能力,因而它可应用于数据挖掘、信息处理、图像识别等多个方面。

Microsoft 神经网络支持 Microsoft 决策树可以执行的所有任务,包括分类、回归和关联。前两个任务是神经网络最常见的任务,而关联任务可能太耗时和耗资源,所以一般不推荐使用神经网络。Microsoft 神经网络算法在使用中有可以调整的参数,均可以根据需要及挖掘结果进行调整。在本章案例中,使用了参数的默认值。

7.3 具体实现

在进行数据挖掘之前，需要建立“数据源”和“数据源视图”。本章的案例为“某商业银行信贷”数据库，其中记录了客户的背景数据以及贷款情况，银行需要从这些数据中发现客户背景与不良贷款记录之间的关系，即发现具有哪些背景的用户更容易产生不良贷款，从而对贷款去向进行监督，避免损失，也可以帮助银行为优质客户提供更好的服务。

7.3.1 数据准备

图 7-2 和图 7-3 所示分别是“某商业银行信贷”数据中的两个基本表：客户基本情况表和贷款余额表。

表 - dbo.客户基本情况表 | 摘要

客户代码	客户名称	客户类型	经济性质	隶属关系	法人资格	客户状态	重点标志
77020101000060	K060单位	工业	其他股份制	省属	法人	正常	一级重点
77020105000009	K009单位	其他	其他	地州市属	法人	正常	非重点
77020105000010	K010单位	工业	国有	省属	法人	正常	一级重点
77020105000013	K013单位	其他	其他	省属	法人	正常	非重点
77020105000025	K025单位	商业	国有控股	地州市属	法人	正常	总行重点
77020105000037	K037单位	商业	其他	省属	法人	停产	非重点
77020105000040	K040单位	商业	民营	其他	法人	停产	非重点
77020105000042	K042单位	工业	集体	地州市属	法人	正常	非重点
77020105000104	K104单位	工业	国有	中央	法人	正常	一级重点
77020105000112	K112单位	其他	其他	其他	法人	停产	非重点
77020105000120	K120单位	工业	其他	无隶属	法人	停产	非重点
77020105000121	K121单位	其他	其他	无隶属	法人	停产	非重点
77020105000122	K122单位	商业	外贸	其他	二级法人	停产	非重点
77020105000136	K136单位	工业	国有控股	中央	法人	正常	非重点
77020105000139	K139单位	工业	国有控股	中央	法人	正常	一级重点
77020105000145	K145单位	商业	外贸	省属	法人	正常	非重点
77020105000157	K157单位	商业	私营	无隶属	法人	正常	非重点
77020105000158	K158单位	商业	外贸	其他	法人	正常	非重点
77020108000001	K001单位	工业	国有控股	省属	授权法人	正常	一级重点
77020108000016	K016单位	工业	三资	无隶属	法人	正常	非重点
77020108000045	K045单位	供销	国有	省属	授权法人	正常	非重点
77020108000049	K049单位	商业	国有控股	地州市属	法人	半停产	非重点
77020108000050	K050单位	商业	国有	中央	法人	正常	一级重点

图 7-2 客户基本情况数据

表 - dbo.贷款余额表 | 表 - dbo.客户基本情况表 | 摘要

分行代码	客户代码	业务发生日	余额	正常	关注	次级	可疑	损失
7702	77020101000060	2002-3-20 0:00:00	44000000.00	44000000.00	0.00	0.00	0.00	0.00
7702	77020105000009	2002-8-30 0:00:00	2700000.00	0.00	0.00	0.00	2700000.00	0.00
7702	77020105000009	2002-8-30 0:00:00	1200000.00	0.00	0.00	0.00	1200000.00	0.00
7702	77020105000009	2002-7-30 0:00:00	3000000.00	0.00	0.00	0.00	3000000.00	0.00
7702	77020105000009	2002-9-25 0:00:00	500000.00	0.00	0.00	0.00	500000.00	0.00
7702	77020105000009	2002-9-25 0:00:00	1410000.00	0.00	0.00	0.00	1410000.00	0.00
7702	77020105000009	2002-8-30 0:00:00	1400000.00	0.00	0.00	0.00	1400000.00	0.00
7702	77020105000010	2002-8-31 0:00:00	526457.00	526457.69	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	10474.00	10474.58	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	70623.00	70623.69	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	20393.00	20393.01	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	468.00	468.83	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	20882.00	20882.30	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	3519.00	3519.27	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	298505.00	298505.28	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	54395.00	54395.64	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	83952.00	83952.70	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	6266.00	6266.48	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	313747.00	313747.75	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	1271511.00	1271511.10	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	450770.00	450770.20	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	560833.00	560833.51	0.00	0.00	0.00	0.00
7702	77020105000010	2002-8-31 0:00:00	21508.00	21508.80	0.00	0.00	0.00	0.00

图 7-3 贷款余额数据

首先为了区分贷款余额表中的每一项业务，在该表中添加了一个主键列，命名为“业务号”，使用“smallint”数据类型。

然后，为了简化挖掘过程，将这两张表的信息合并到同一张表中，合并方法如图 7-4 所示的 SQL 语句。合并后产生的新表命名为 t_dm。

```
2B0E6E3327A1...QLQuery1.sql*  表 - dbo.贷款余额表  表 - dbo.客户基本情况表  摘要
select a.*,b.正常,b.关注,b.次级,b.可疑,b.损失,b.业务号
into t dm
from 客户基本情况表 as a
join 贷款余额表 as b
on a.客户代码=b.客户代码
```

图 7-4　用于合并的 SQL 数据

与其他挖掘模型一样，神经网络模型需要做一些设置，包括两个方面：指定列的用法和设置挖掘参数。其中，挖掘参数设置本实例使用默认设置。列的用法如表 7-1 所示。

表 7-1　列的用法

字段名称	列的用法	字段名称	列的用法	字段名称	列的用法	字段名称	列的用法
ID	键列	隶属关系	输入列	重点标志	输入列	余额	预测列
客户名称	输入列	关注	输入列	可疑	预测列	损失	预测列
客户类型	输入列	法人资格	输入列	次级	预测列		
经济性质	输入列	客户状态	输入列	正常	预测列		

为了满足神经网络对数据的要求，即神经网络要求输入数据都是数值类型，对该数据库进行数据类型转换处理，将 t_dm 表中用作输入列的文本类型数据转换成数值类型。下面以输入列“经济性质”为例，来说明具体转换过程，其他列的转换也是类似的。

（1）统计输入列中不同类别的个数，SQL 实现语句及执行结果如图 7-5 所示。

```
select count(*),经济性质from t_dm group by经济性质
```

(a) 统计类别SQL语句

结果　消息

	(无列名)	经济性质
1	187	个体
2	1	研究所
3	408	股份合作
4	36	民营
5	208	其他
6	72	三资
7	1	部队
8	18	学校
9	1262	国有
10	419	集体
11	1	联营
12	55	其他股份制
13	13	外贸
14	15	医院
15	38	机关团体
16	496	国有控股
17	106	私营
18	74	集体控股

(b) 执行结果

图 7-5　位计类别

（2）根据步骤（1）统计信息用数值来替代文本数据，SQL 实现语句及执行结果如图 7-6 所示。

```
update t_dm set 经济性质='1' where 经济性质='国有'
update t_dm set 经济性质='2' where 经济性质='其他股份制'
update t_dm set 经济性质='3' where 经济性质='其他'
update t_dm set 经济性质='4' where 经济性质='国有控股'
update t_dm set 经济性质='5' where 经济性质='民营'
update t_dm set 经济性质='6' where 经济性质='集体'
update t_dm set 经济性质='7' where 经济性质='外贸'
update t_dm set 经济性质='8' where 经济性质='私营'
update t_dm set 经济性质='9' where 经济性质='三资'
update t_dm set 经济性质='10' where 经济性质='股份合作'
update t_dm set 经济性质='11' where 经济性质='集体控股'
update t_dm set 经济性质='12' where 经济性质='研究所'
update t_dm set 经济性质='13' where 经济性质='机关团体'
update t_dm set 经济性质='14' where 经济性质='医院'
update t_dm set 经济性质='15' where 经济性质='学校'
update t_dm set 经济性质='16' where 经济性质='个体'
update t_dm set 经济性质='17' where 经济性质='部队'
update t_dm set 经济性质='18' where 经济性质='联营'
```

（a）更新表中数据 SQL 语句

```
消息
(1262 行受影响)
(55 行受影响)
(208 行受影响)
(496 行受影响)
(36 行受影响)
(419 行受影响)
(13 行受影响)
(106 行受影响)
(72 行受影响)
(408 行受影响)
(74 行受影响)
(1 行受影响)
(38 行受影响)
(15 行受影响)
(18 行受影响)
(187 行受影响)
(1 行受影响)
```

(b) 执行结果

图 7-6　数据替代命令及结果

所有输入列经过以上两步处理之后，部分 t_dm 表中数据如图 7-7 所示。

表 - dbo.t_dm　new.bank - SQLQuery1.sql*　表 - dbo.t_bank　摘要

客户名称	客户类型	经济性质	隶属关系	法人资格	客户状态	重点标志
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K010单位	1	1	8	2	1	1
K013单位	8	3	8	2	1	4
K025单位	2	4	6	2	1	5
K037单位	2	3	8	2	2	4
K037单位	2	3	8	2	2	4
K040单位	2	5	3	2	2	4
K040单位	2	5	3	2	2	4
K040单位	2	5	3	2	2	4
K040单位	2	5	3	2	2	4
K040单位	2	5	3	2	2	4
K040单位	2	5	3	2	2	4
K042单位	1	6	6	2	1	4
K042单位	1	6	6	2	1	4
K104单位	1	1	10	2	1	1
K112单位	8	3	3	2	2	4
K120单位	1	3	5	2	2	4
K121单位	8	3	5	2	2	4
K121单位	8	3	5	2	2	4

图 7-7　处理后数据

为了得到更简洁挖掘结果，对该数据库中“t_dm”表的次级、可疑、损失、余额和正常5个字段进行了更新处理，SQL处理语句如下：

```
Update 贷款余额表 set 次级=1 where 次级> 0.0
Update 贷款余额表 set 可疑=1 where 可疑> 0.0
Update 贷款余额表 set 损失=1 where 损失> 0.0
Update 贷款余额表 set 余额=1 where 余额> 0.0
Update 贷款余额表 set 正常=1 where 正常> 0.0
```

7.3.2 挖掘流程

（1）从Windows菜单启动Microsoft Visual Studio，如图7-8所示。

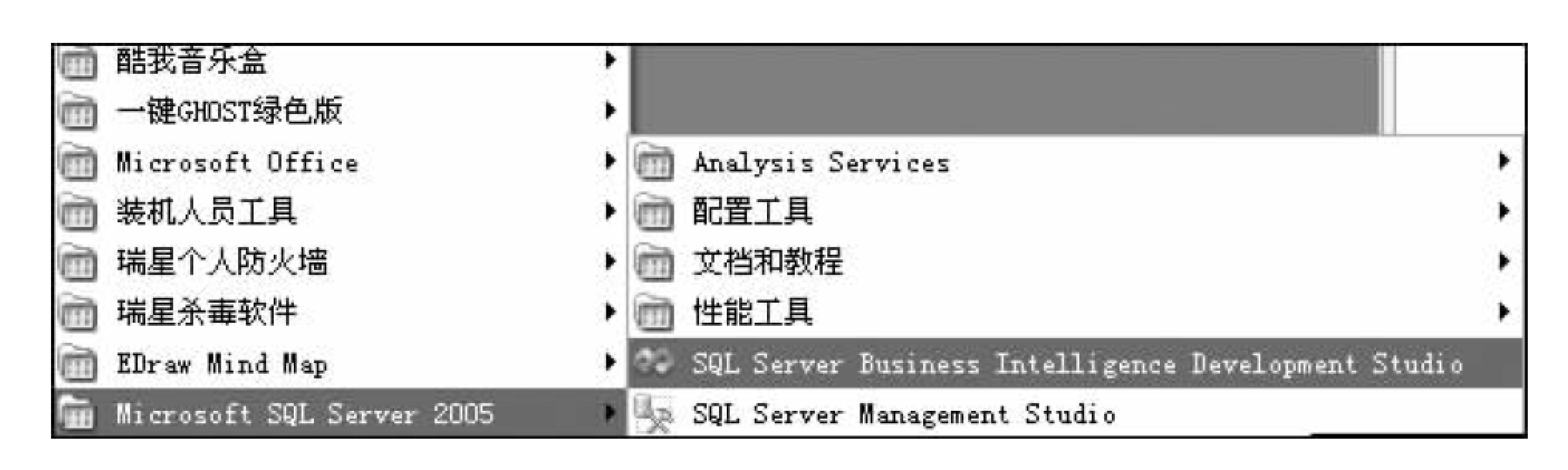


图7-8 启动Microsoft Visual Studio

（2）关闭起始页，选择“文件”→“新建”→“项目”命令，打开“新建项目”对话框。新建一个Analysis Services项目，并且在对话框中指定项目名称和存放位置，如图7-9所示。



图7-9 新建项目

（3）打开解决方案资源管理器，查看已经创建的解决方案，右击“数据源”项，在弹出的菜单中选择“新建数据源”命令，并在数据源向导界面单击“下一步”按钮，如图7-10和图7-11所示。



图 7-10　选择新建数据源　　　　图 7-11　使用数据源向导

(4) 进入"选择如何定义连接",选择"基于现有连接或新连接创建数据源"单选按钮,单击"新建"按钮,如图 7-12 所示。



图 7-12　定义连接

(5) 在打开的"连接管理器"对话框中,选择"本机 OLE DB\SQL Native Client"项,选择服务器名为本机,使用 Windows 身份验证登录,并选择数据库名称 bank,单击"确定"按钮,如图 7-13 所示。

(6) 返回数据源向导,选中已经建立连接的数据库,单击"下一步"按钮,如图 7-14 所示。

(7) 选择分析服务器使用"使用服务账户"单选按钮,作为连接数据源的凭证,并单击

图 7-13　设置连接管理器



图 7-14　返回数据源向导

"下一步"按钮,如图 7-15 所示。

(8) 完成数据源的创建,如图 7-16 所示。

(9) 建立数据源视图。数据源视图提供一组已经存在、可浏览、持久化数据库对象(如表、视图和关系)。Analysis Services 中的联机分析处理(OLAP)和数据挖掘对象可以引用这些数据库对象。可以对这些对象进行组织和配置,以便为数据源提供完整的架构表示

数据源向导

模拟信息

可以定义 Analysis Services 使用何种凭据来连接到数据源。

使用特定用户名和密码(S)

用户名(U):

密码(P):

使用服务帐户(V)

使用当前用户的凭据(D)

默认值(F)

< 上一步(B)　下一步(N) >　完成(F) >>|　取消

图 7-15　设置模拟信息



图 7-16　完成数据源创建

形式。在 Analysis Services 项目或部署数据库中生成数据源视图后，该数据源视图就可供 Analysis Services 中的任何 OLAP 或数据挖掘对象使用。创建数据源视图的方法同创建数据源相同，使用资源管理器中的右键菜单，如图 7-17 所示。

(10) 单击“数据源视图向导”对话框的“下一步”按钮，为数据源视图选择数据源，如图 7-18 所示。

(11) 为数据源视图选择表或视图，这里选择在数据准备工作中建立的表 t_dm，如图 7-19 所示。

(12) 完成数据源视图建立，并命名，单击“完成”按钮，如图 7-20 所示。

(13) 回到 Visual Studio 主界面，右击项目 Neutral Network 下的“挖掘结构”选择“新建挖掘结构”项，打开“数据挖掘向导”对话框，单击“下一步”按钮，切换到“选择定义方法”页面，单击“下一步”按钮，切换到“选择数据挖掘技术”页面。



图 7-17　选择新建数据源视图



图 7-18　选择数据源



图 7-19　选择表和视图



图 7-20　完成数据源视图建立

（14）如图 7-21 所示，在下拉列表框中选取“Microsoft 神经网络”选项，单击“下一步”按钮，切换到下一个页面。



图 7-21　选择神经网络挖掘技术

（15）如图 7-22 所示，在“选择数据源视图”页面的“可用数据源视图”列表中显示了前面步骤创建的 bank 数据源视图，选中该视图选项，单击“下一步”按钮，切换到下一个页面。



图 7-22　选择数据源视图

（16）如图 7-23 所示，在“指定表类型”页面中可以看到 bank 数据源视图包含的数据表，选中 t_dm 选项右边的“事例”复选框，可以将其定义为事例表；单击“下一步”按钮切换到下一个页面。

（17）如图 7-24 所示，在“指定定型数据”页面显示了挖掘模型结构，在各个选项右边选中不同的复选框，可参照表 7-1 完成，然后单击“下一步”按钮，切换到下一个界面。

（18）如图 7-25 所示，经过“检测”将指定数字列，即“次级”、“关注”、“可疑”、“损失”、“余额”和“正常”的连续值转换成离散值，即 0 或 1，与数据处理结果对应起来。在“指定列的内容和数据类型”页面中显示了指定 ID 的内容类型为 Key，“余额”的内容类型为 Continuous，其余



图 7-23　指定表类型

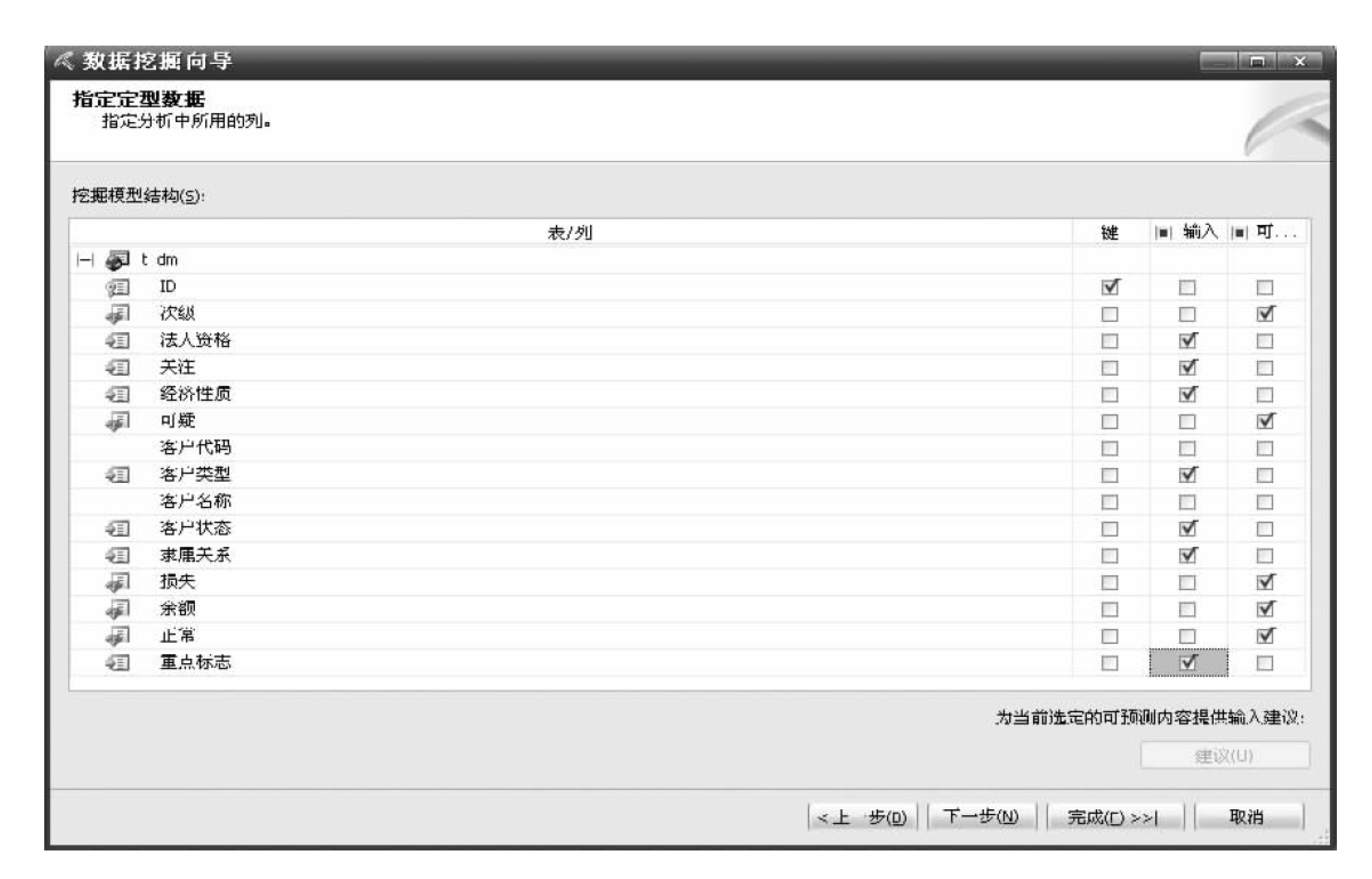


图 7-24　指定定型数据

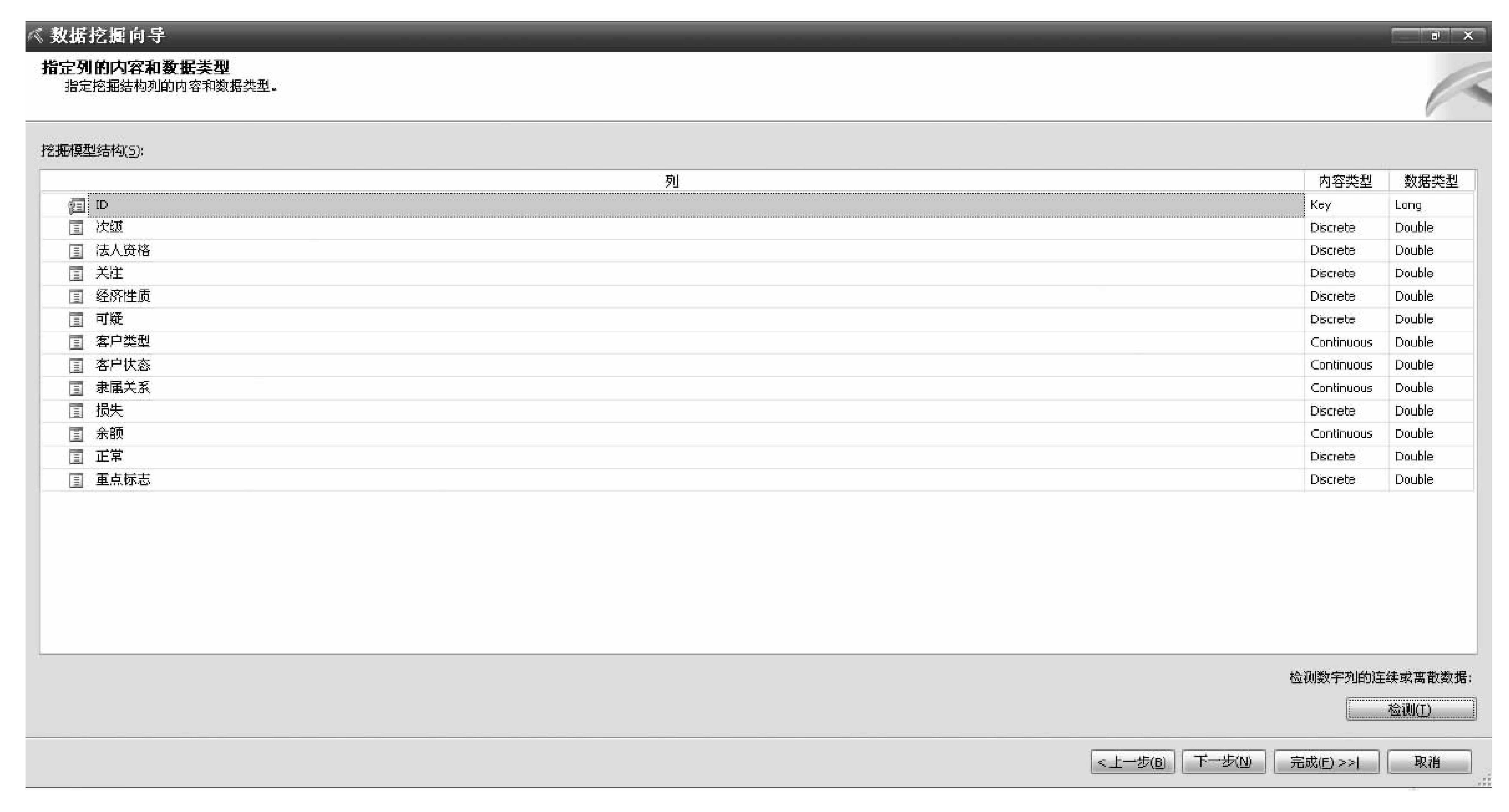


图 7-25　指定列的内容和数据类型

列内容类型均为 Discrete;ID 的数据类型为 long,其余各列数据类型均为 Double,单击“下一步”按钮切换到下一页面。

(19) 如图 7-26 所示,在“完成向导”页面中将数据挖掘结构命名为 t_Dm1,单击“完成”按钮,完成挖掘结构的创建。



图 7-26　完成挖掘结构创建

(20) 单击“挖掘准确性图表”选项卡下的“提升图”和“分类矩阵”,其结果如图 7-27 和图 7-28 所示。

图 7-27　查看提升图

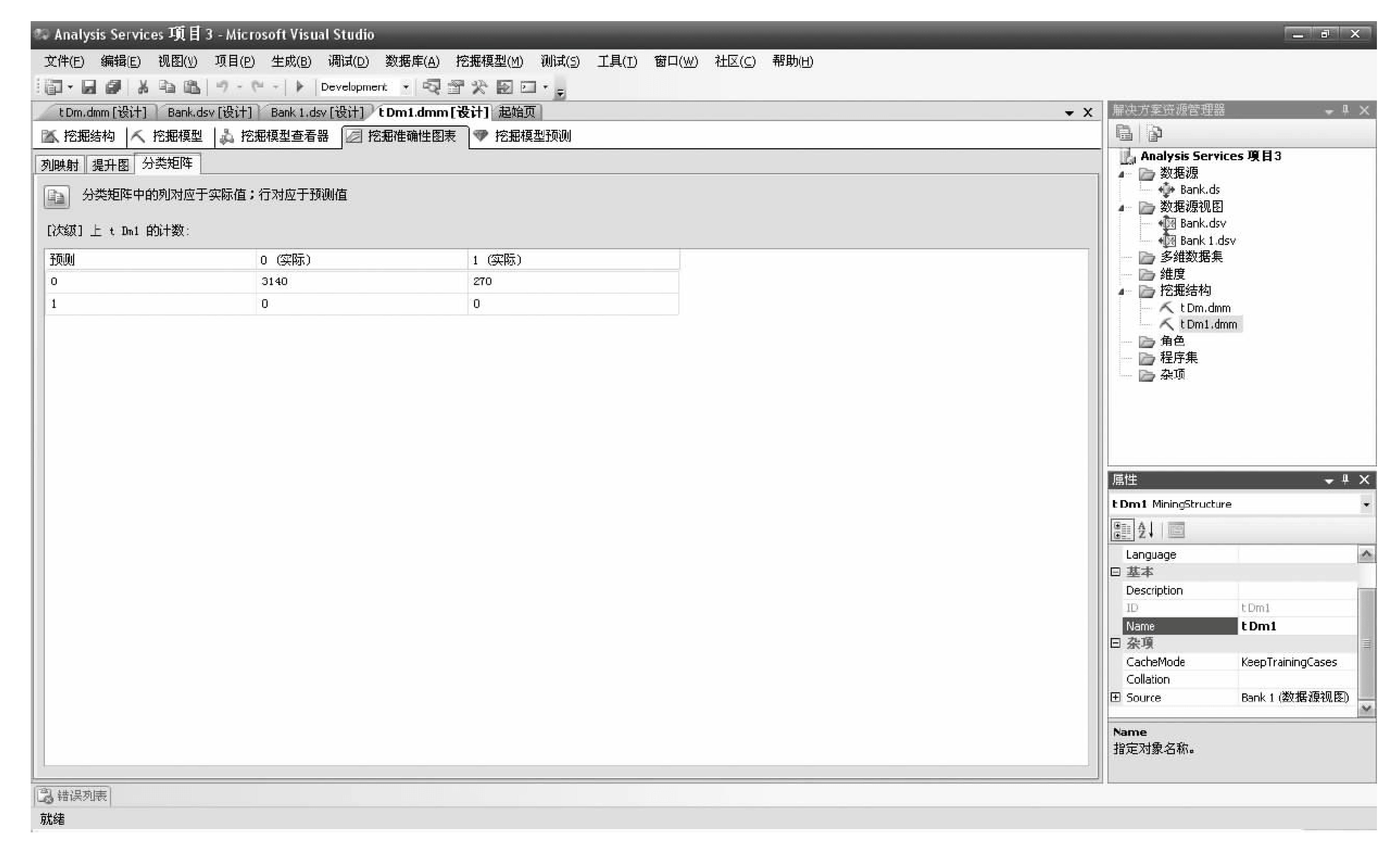


图 7-28　查看分类矩阵

在图 7-27 中的下方线代表神经网络建立的预测模型，上方线代表实际模型，可以看出两者是基本匹配的，说明预测模型比较理想；在图 7-28 中，对“次级”=1 的预测结果是无差错的，对“次级”=0 的预测结果有 7.9%的误差。

7.4　案例小结

在本章案例中，首先针对性地获得了电子数据，然后通过事实数据表的有效数据项的选取构建了待挖掘数据集，根据现实情况，选取了与信贷分级关系密切的数据项：“客户名称”、“客户类型”、“经济性质”、“隶属关系”、“关注”、“法人资格”、“客户状态”和“重点标志”作为输入列，将信贷 5 个等级作为输出列，通过 Microsoft 神经网络模型分析输入与输出的关联，对现有数据进行了等级分析。这是整个挖掘的思路和过程。

特别指出，利用神经网络模型进行挖掘时，必须对非数值数据进行适当数值化处理。另外，无论使用哪种模型，对原始数据库进行预处理都是必须的，而且往往能影响挖掘结果。

实例8 基于 K-means 方法的栀子花聚类分析

8.1 任务描述

有采集到的150朵栀子花的数据如表8-1所示,每朵栀子花给出了4个属性值,分别为萼片长度(Sepal Length)、萼片宽度(Sepal Width)、花瓣长度(Petal Length)和花瓣宽度(Petal Width)。已知150朵栀子花分属三种不同的类型,用数据挖掘中的K-means方法判断哪些栀子花属于同一类型。

表8-1 栀子花数据

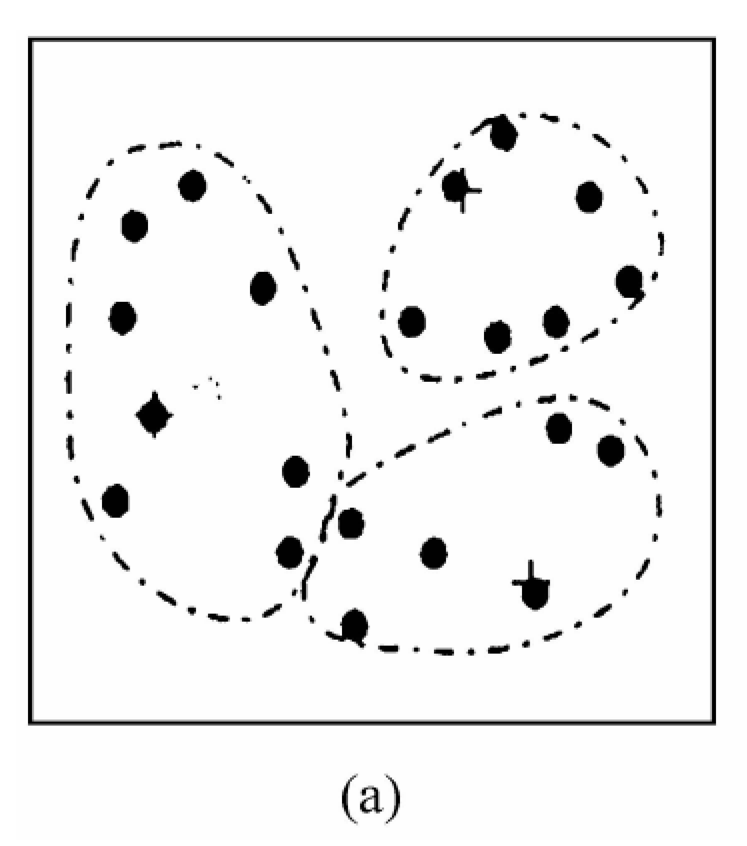
No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3.0	1.4	0.1

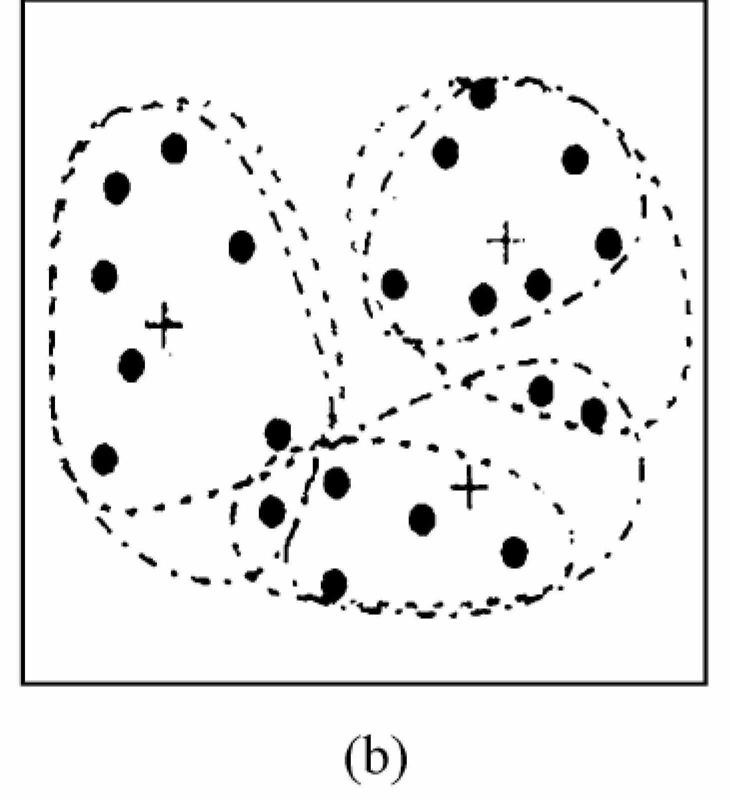
8.2 技术原理

聚类的任务是把所有的实例分配到若干的簇,使得同一个簇的实例聚集在一个簇中心的周围,它们之间的距离比较近;不同簇实例之间的距离比较远。对于由数值型属性刻画的实例来说,这个距离通常指欧氏距离。

聚类分析的方法很多,其中包括基于划分的聚类方法、基于层次的聚类方法、基于密度的聚类方法、基于网格的聚类方法和谱聚类方法等。

K-means方法是一种基于划分的聚类方法。其核心思想是通过迭代把数据对象划分到不同的簇中,以求目标函数最小化,从而使生成的簇尽可能地紧凑和独立。K-means方法的具体划分过程是,首先,随机选取k个对象作为初始的k个簇的质心;然后,将其余对象根据其与各个簇质心的距离分配到最近的簇;最后,再求新形成的簇的质心。如此迭代、重定位,尝试通过对象在划分之间的移动来改进划分。图8-1所示为K-means方法的迭代过程。





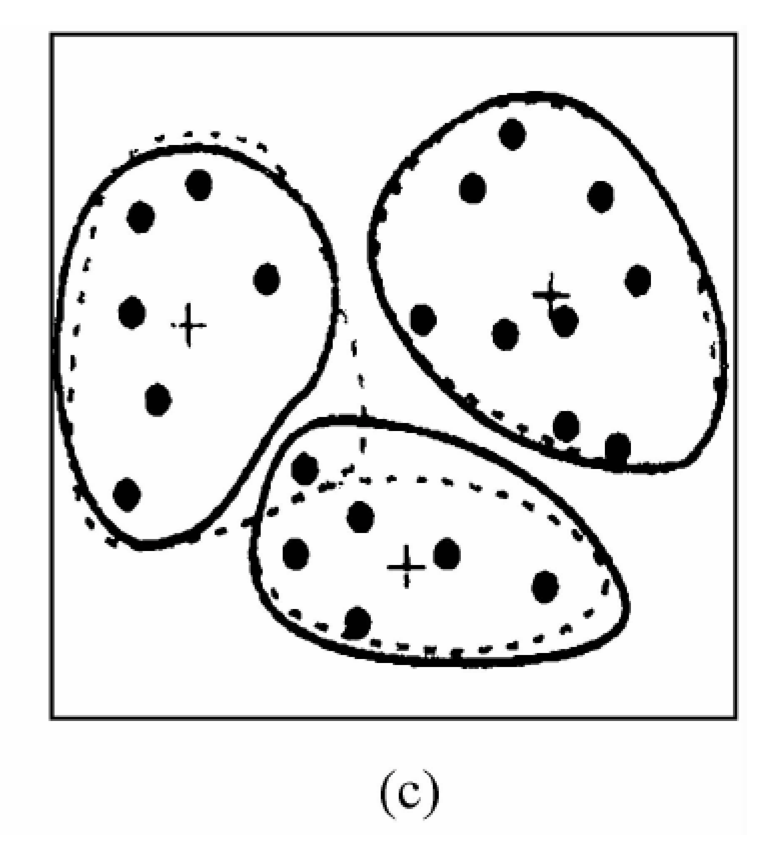


图 8-1　K-means 迭代示例

8.3　具 体 实 现

（1）选择“开始”→“所有程序”→Weka 3.6.5→Weka 3.6 命令，如图 8-2 所示。



图 8-2　打开 Weka 软件

（2）单击 Explorer 按钮，如图 8-3 所示。



图 8-3　打开 Explorer 应用

（3）单击 Open file 按钮，在弹出的对话框中选择要打开的文件 iris.arff，并单击“打开”按钮，如图 8-4 所示。

（4）在如图 8-5 所示的界面中，可以知道 Iris 数据集中共有 150 个实例，每个实例有 5 个属性。选中某个属性，可以查看 150 个实例关于这个属性的属性值的最小值、最大值、均值和标准差等信息。然后单击 Cluster 标签。

（5）单击 Choose 按钮，如图 8-6 所示。



图 8-4　打开数据文件

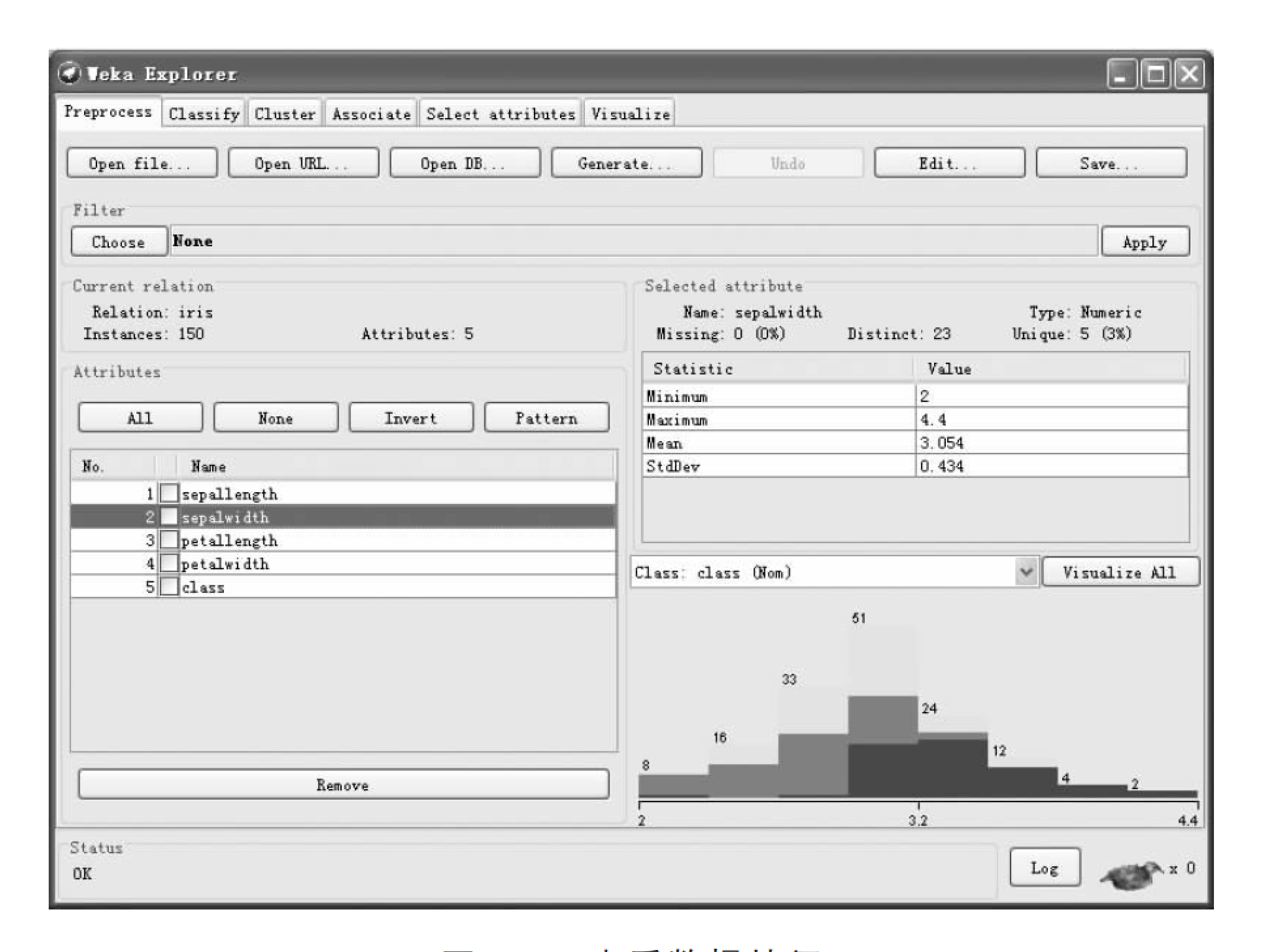


图 8-5　查看数据特征

图 8-6　打开聚类方法选取界面

（6）选择 SimpleKMeans 聚类方法，并单击 Close 按钮，如图 8-7 所示。

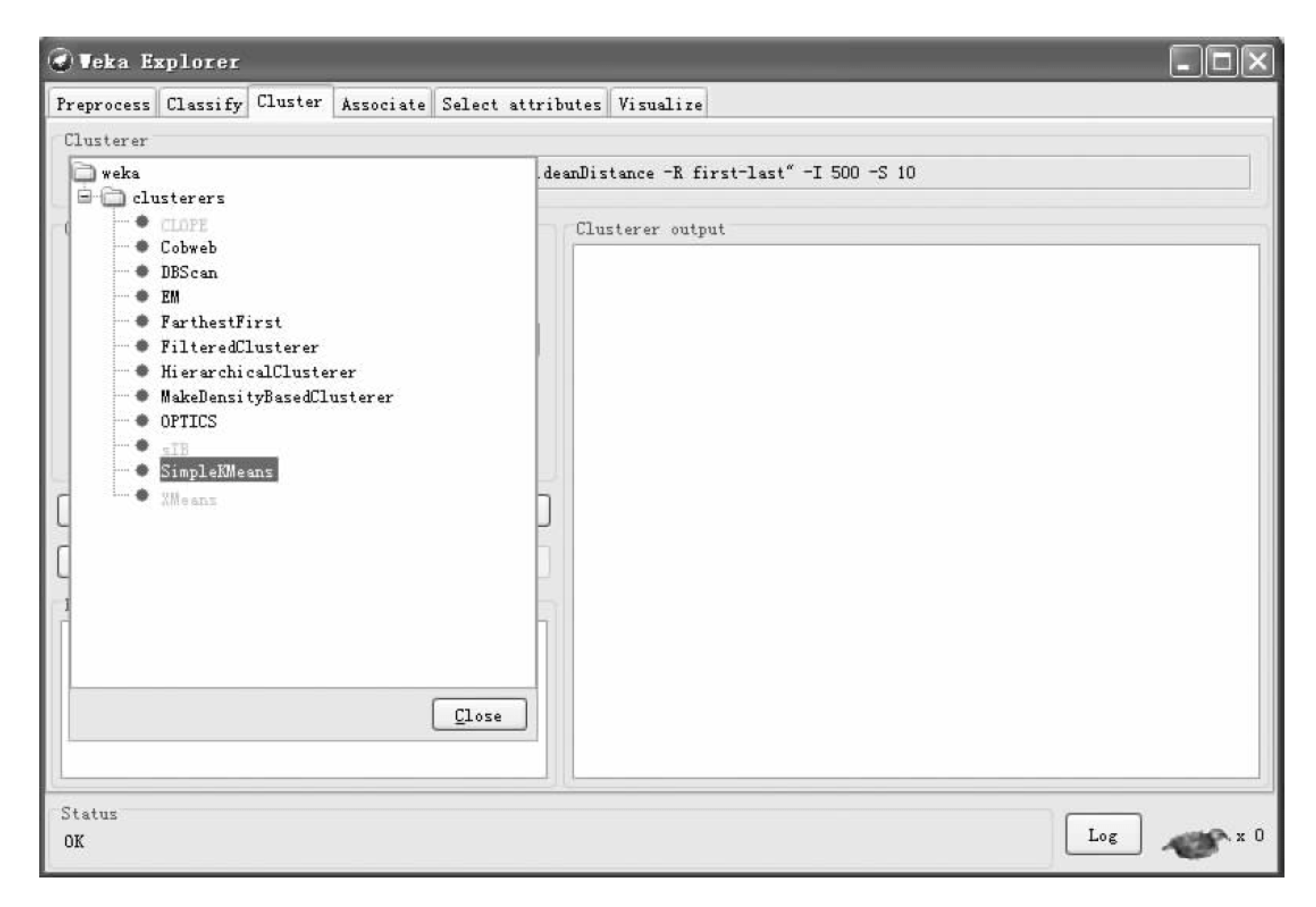


图 8-7　选择聚类方法

（7）单击 Choose 按钮后的 SimpleKMeans 聚类方法，弹出参数设置框，如图 8-8 所示。

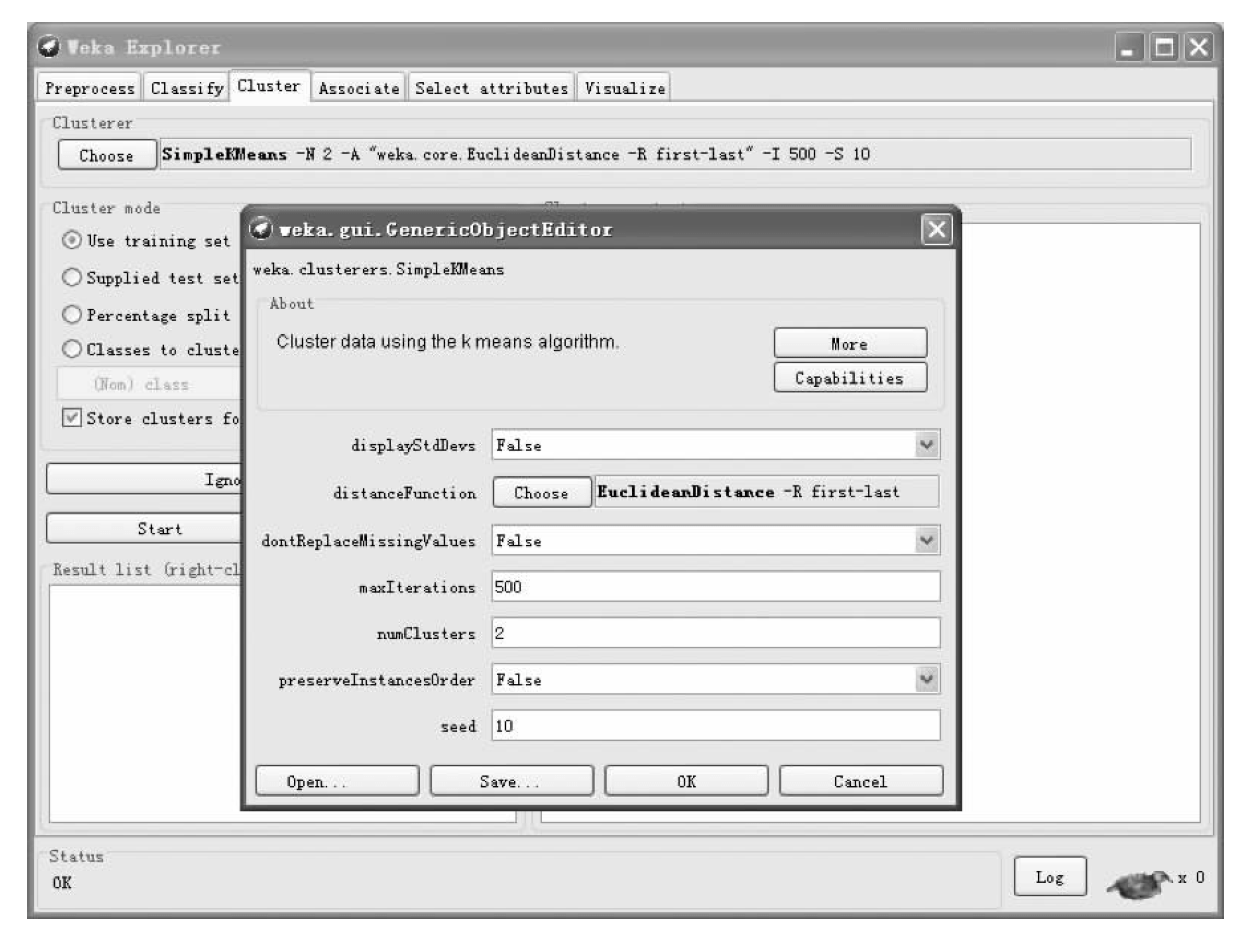


图 8-8　进行参数设置

（8）对聚类方法的参数进行设置，其中 numClusters 设置为 3，并单击 OK 按钮，如图 8-9 所示。

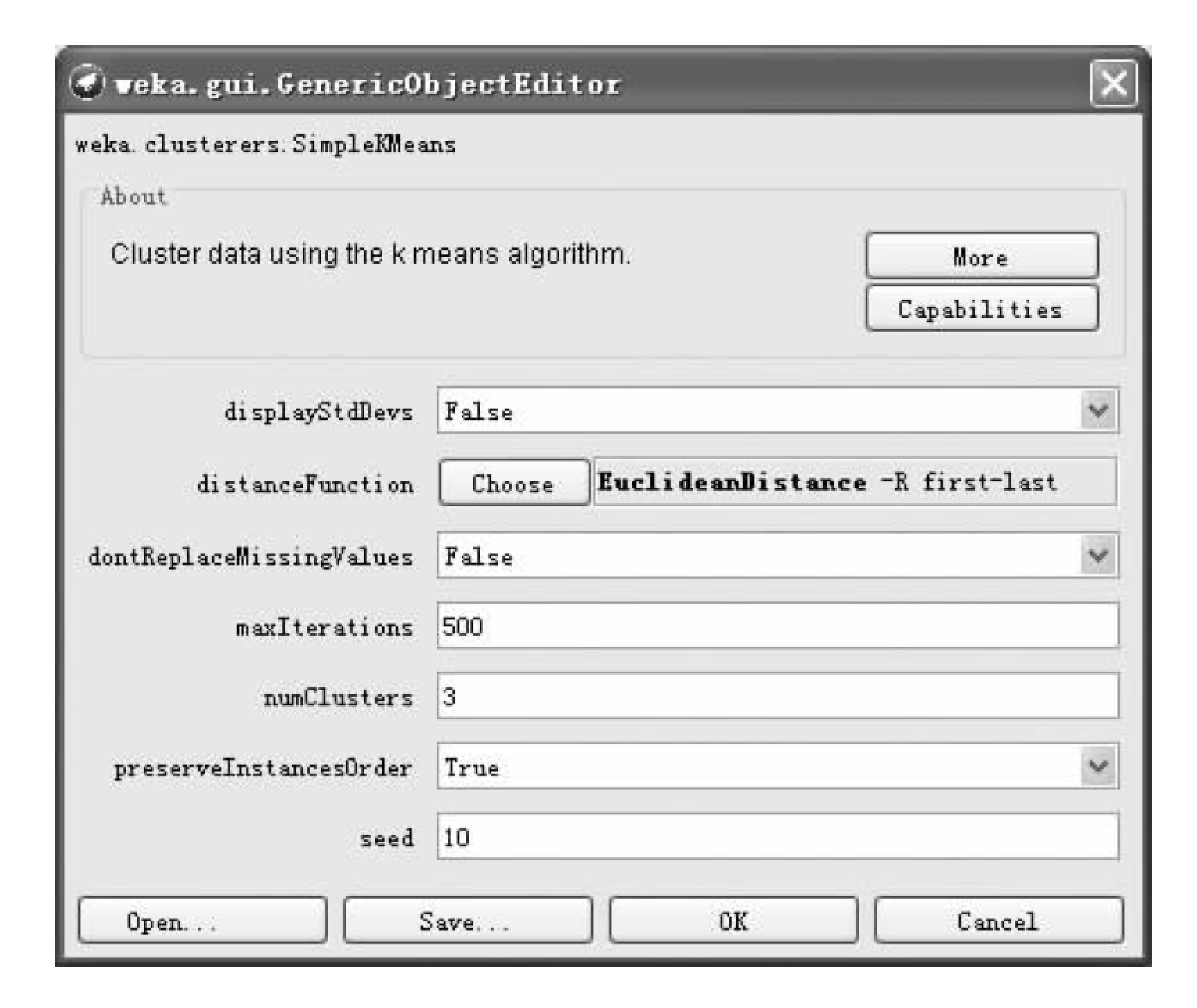


图 8-9　设置簇的个数

(9) 选中 Classes to clusters evaluation 复选框，并将数据集中的 class 属性作为用来评估聚类效果的聚类属性，如图 8-10 所示。

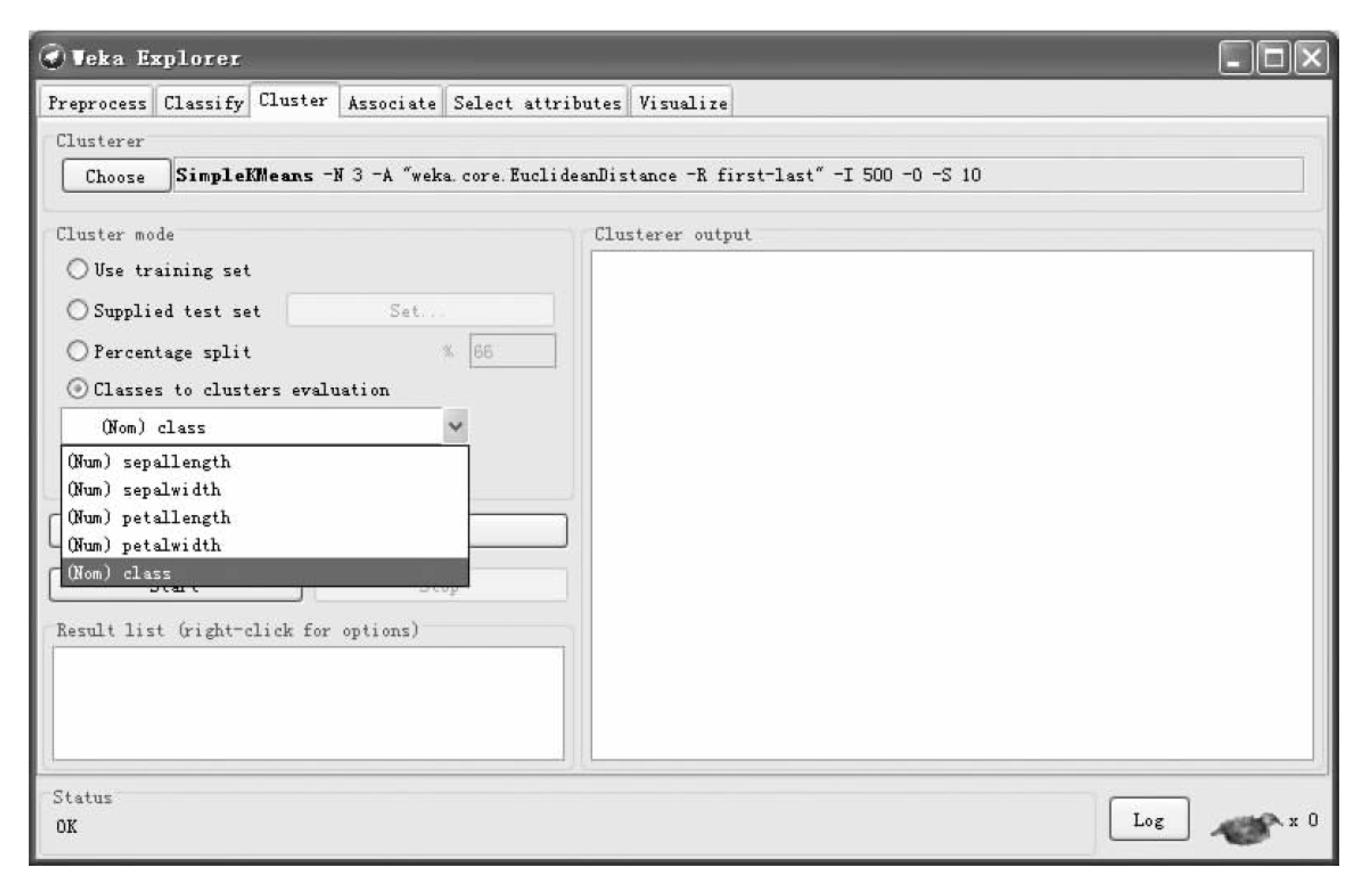


图 8-10　设置评估属性

(10) 单击 Ignore attributes 按钮，在弹出的对话框中选择 class 属性，并单击 Select 按钮，如图 8-11 所示。

(11) 单击 Start 按钮，Weka 对 Iris 数据集执行 KMeans 算法，如图 8-12 所示。

(12) 在算法的执行结果中给出了算法一共迭代了 6 次，同时给出了 3 个簇中心的各个属性值，如图 8-13 所示。

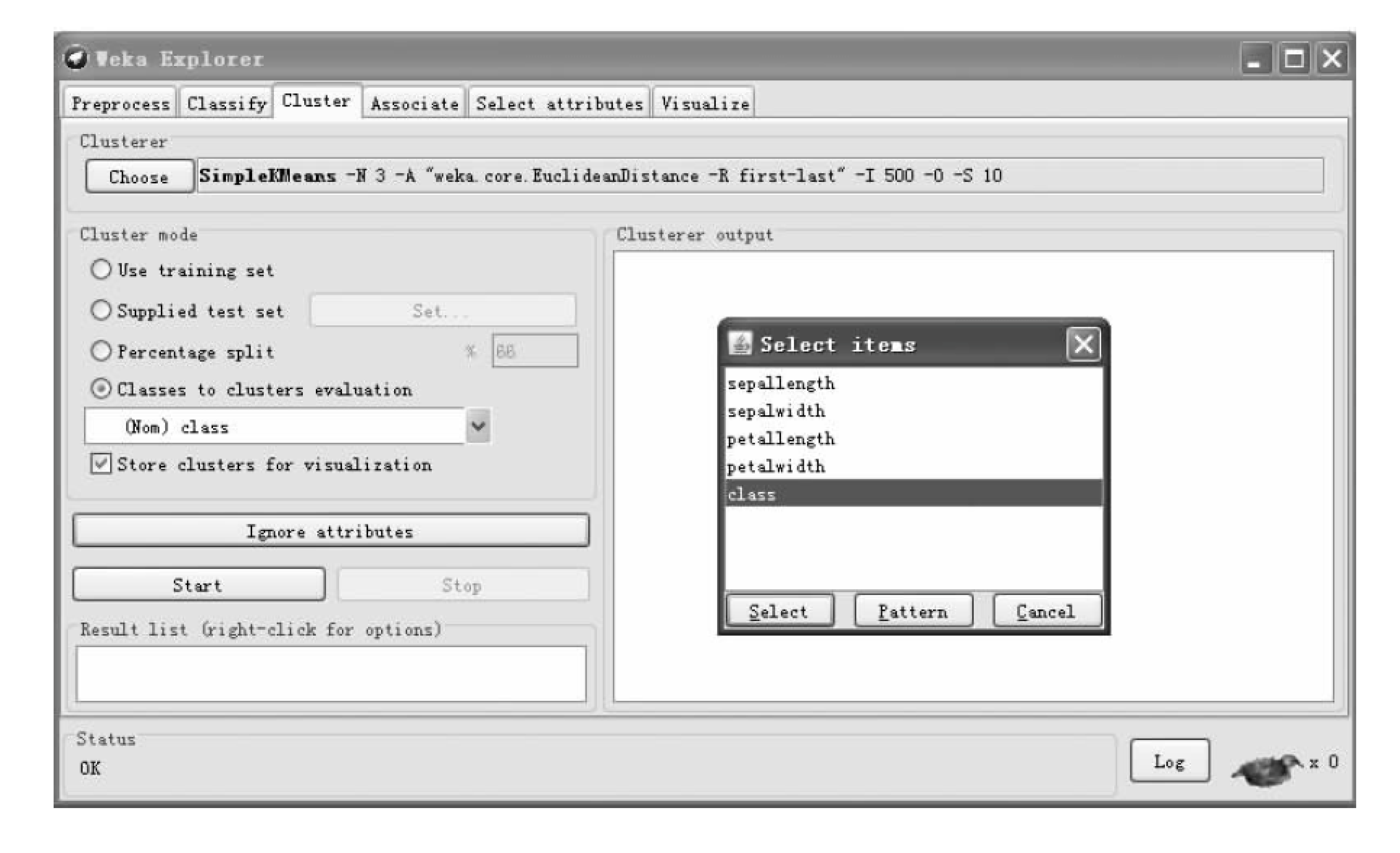


图 8-11 设置忽略属性

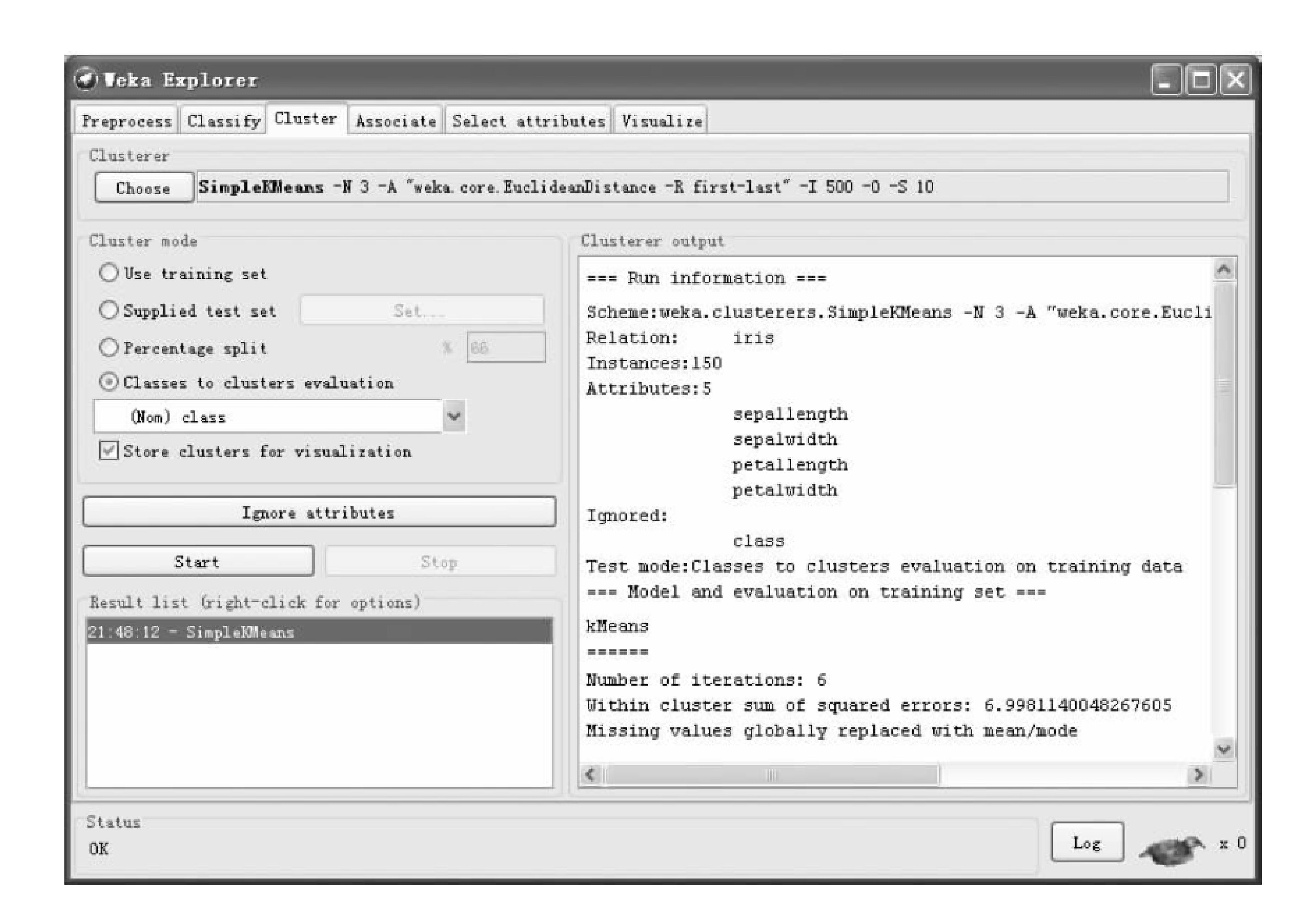


图 8-12 执行聚类算法

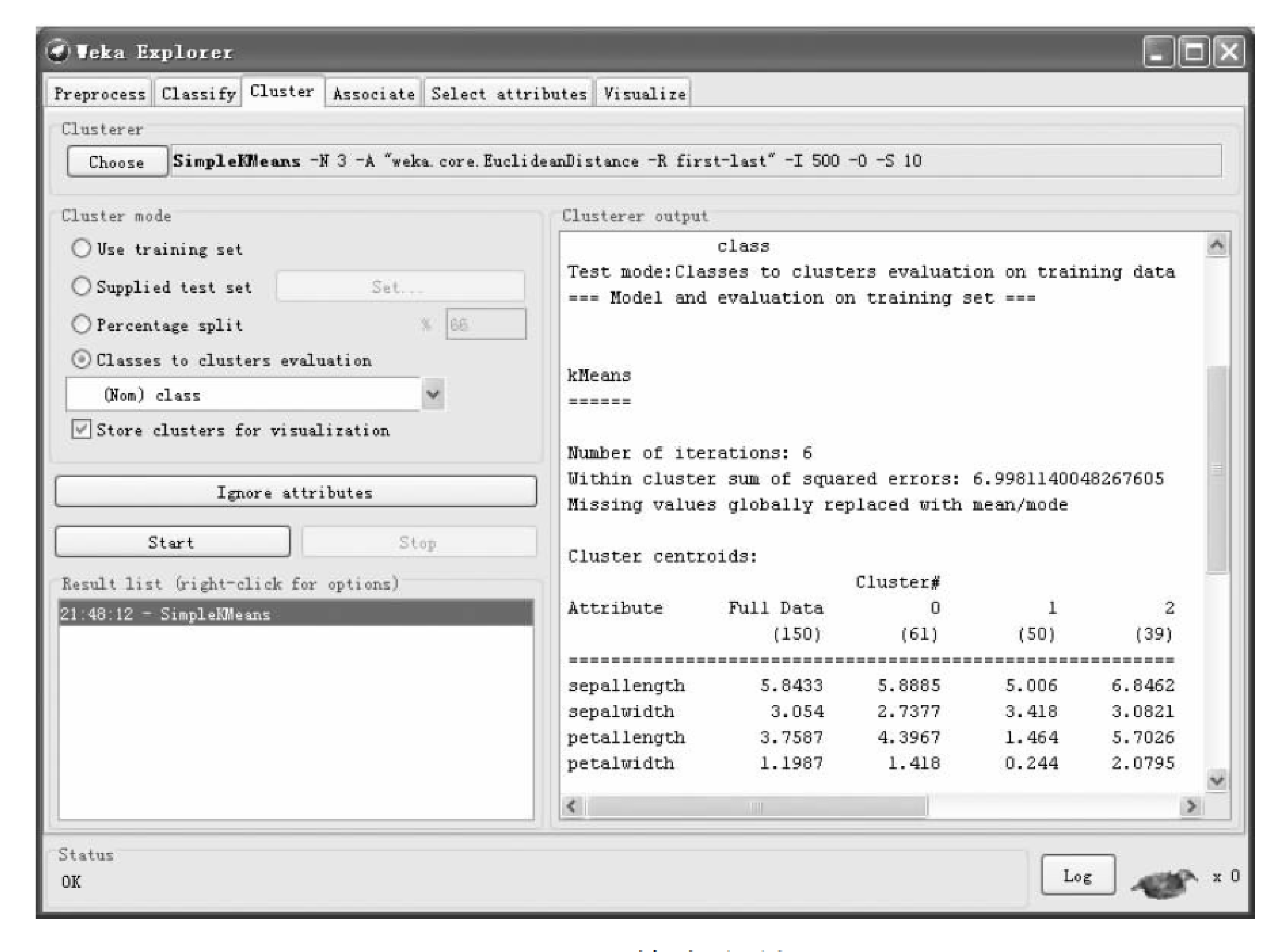


图 8-13 簇中心结果

（13）在算法的执行结果中还给出了每个簇所含实例的个数以及占全体实例的百分比。在已知标准聚类结果的前提下，算法的执行结果还能给出标准簇和通过算法得到的簇之间的对应关系，以及整个聚类结果的错误率，如图 8-14 所示。

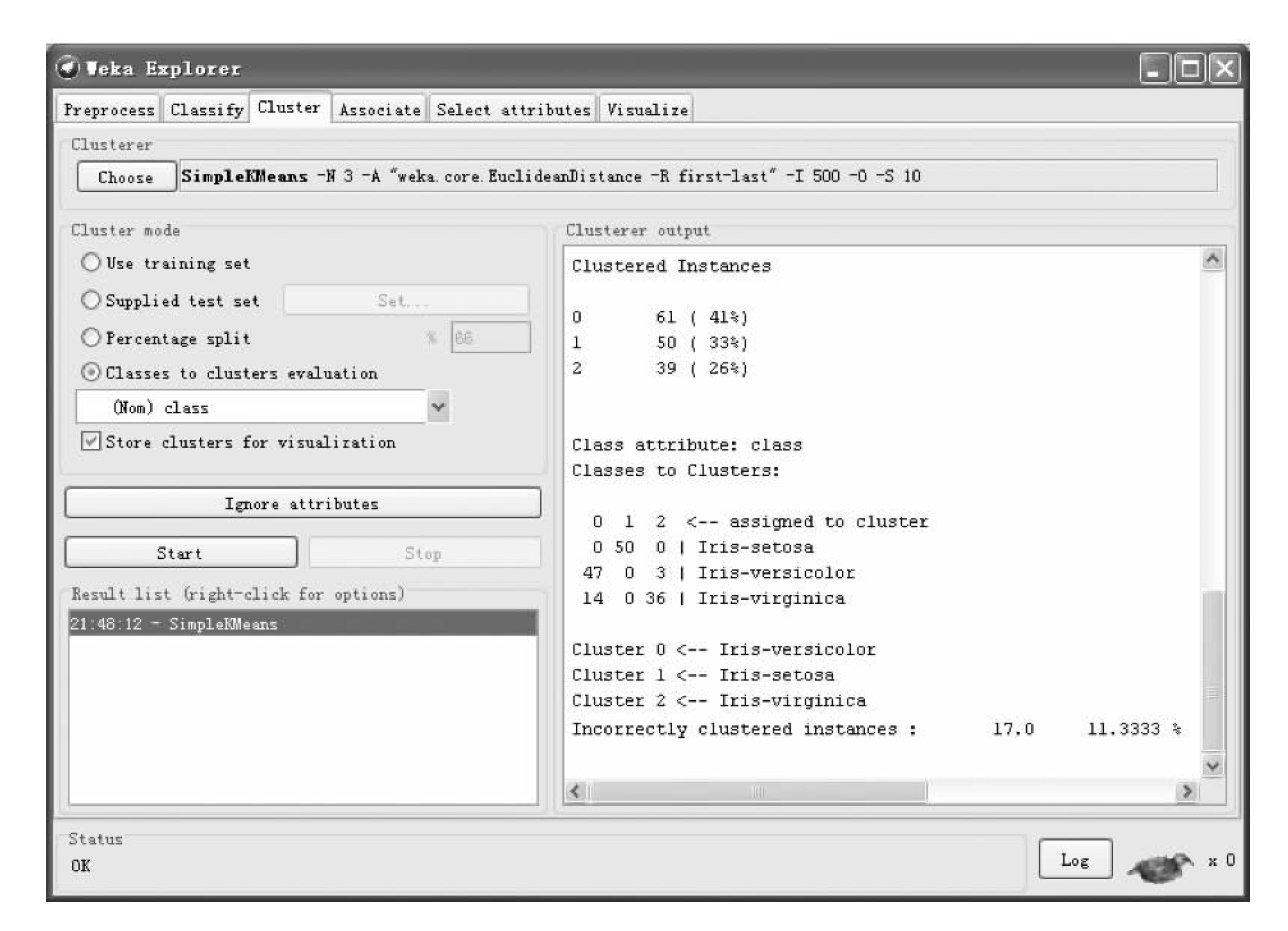


图 8-14　继续进行结果分析

（14）为了观察可视化的聚类结果，在左下方 Result list 列出的结果上右击，在弹出的菜单中选择 Visualize cluster assignments 项，如图 8-15 所示。

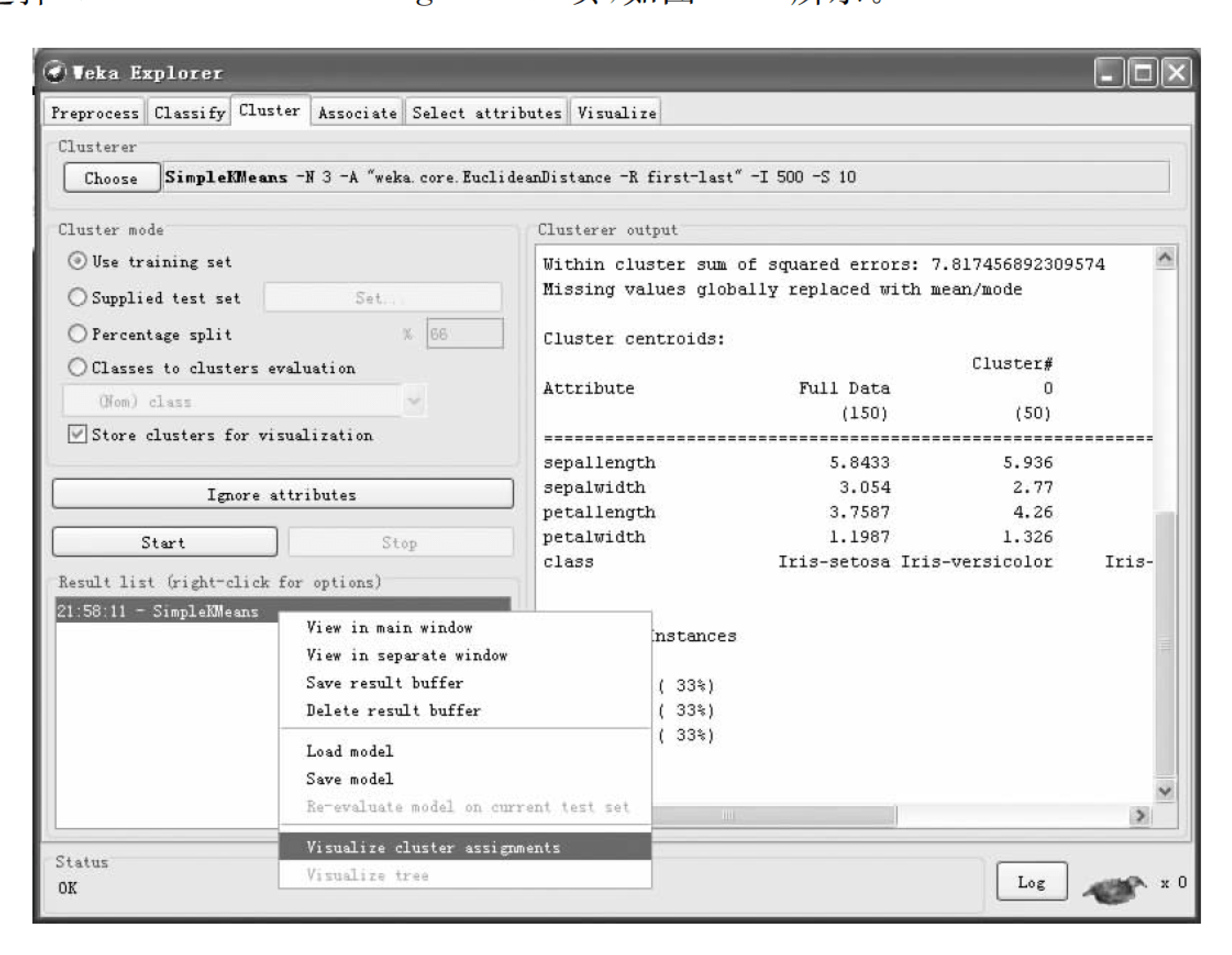


图 8-15　选择可视化聚类结果

(15) 在弹出的可视化结果对话框中，可以查看实例、属性值和簇之间的对应关系，如图 8-16 所示。

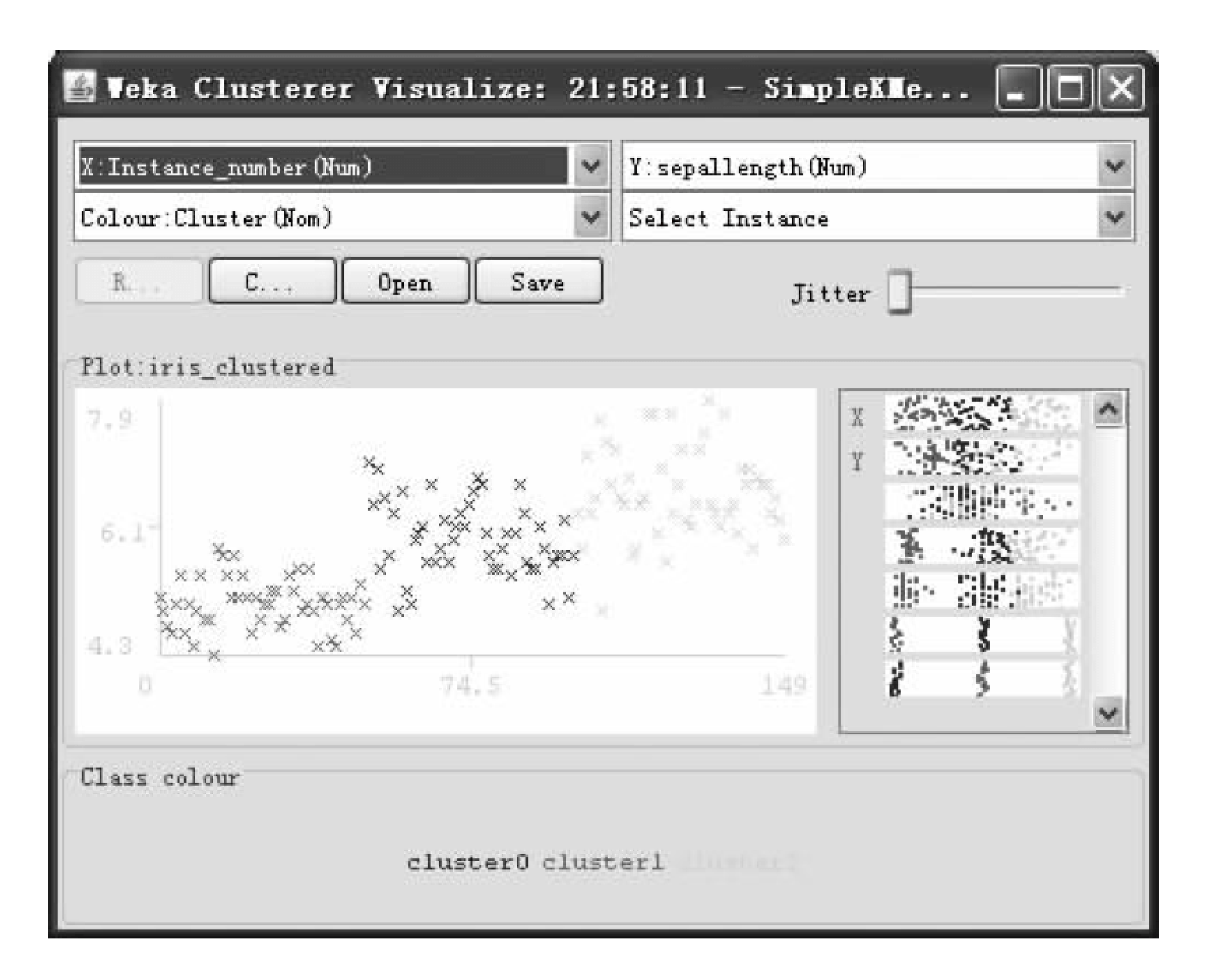


图 8-16　浏览可视化结果

(16) 在可视化结果对话框中，单击 Save 按钮，将结果保存在 irisKMeans. arff 文件中，如图 8-17 所示。

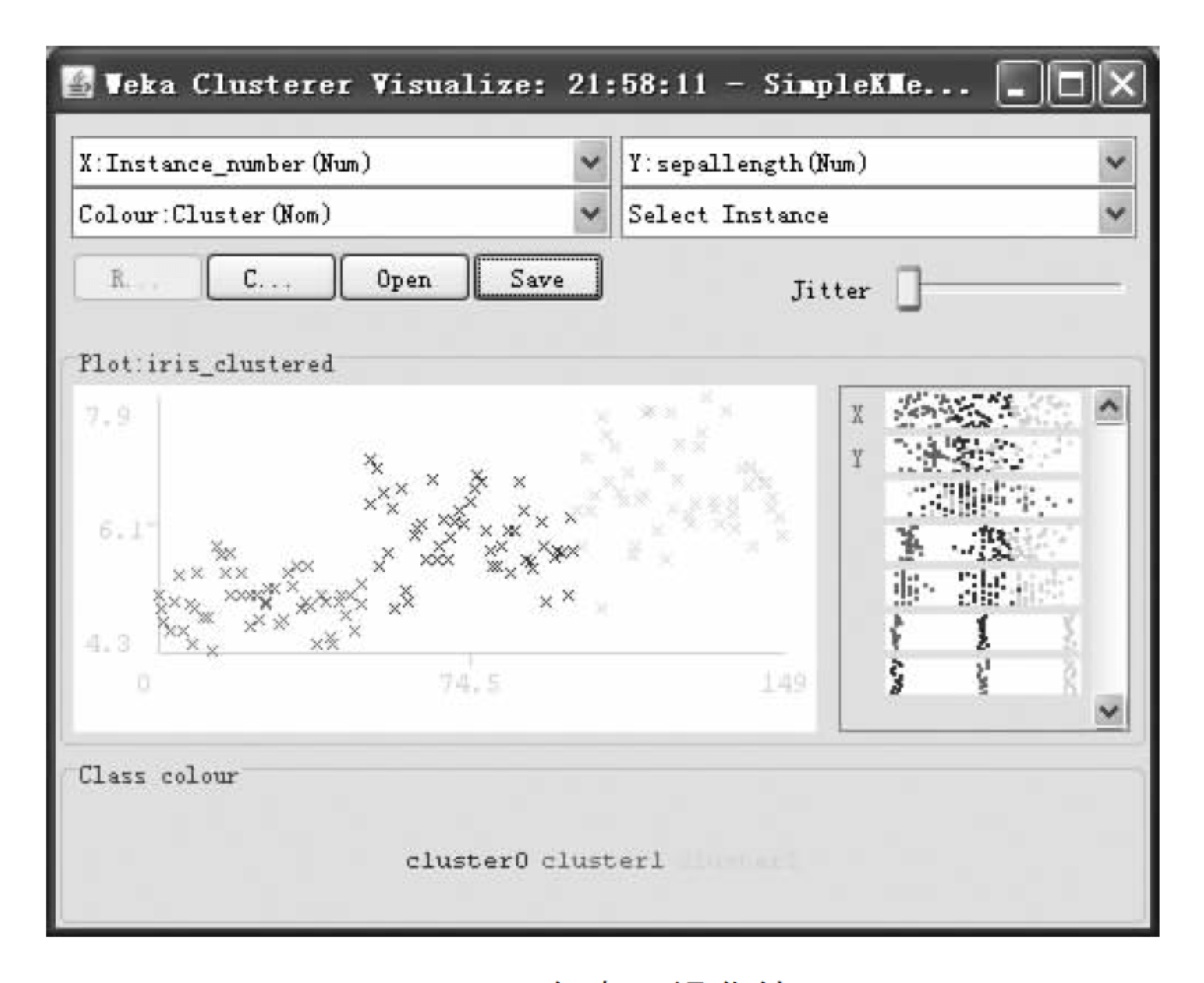


图 8-17　保存可视化结果

(17) 在 Weka Explorer 界面中打开 irisKMeans. arff 文件，instance_number 属性表示某实例的编号，Cluster 属性表示聚类算法给出的该实例所在的簇，如图 8-18 所示。

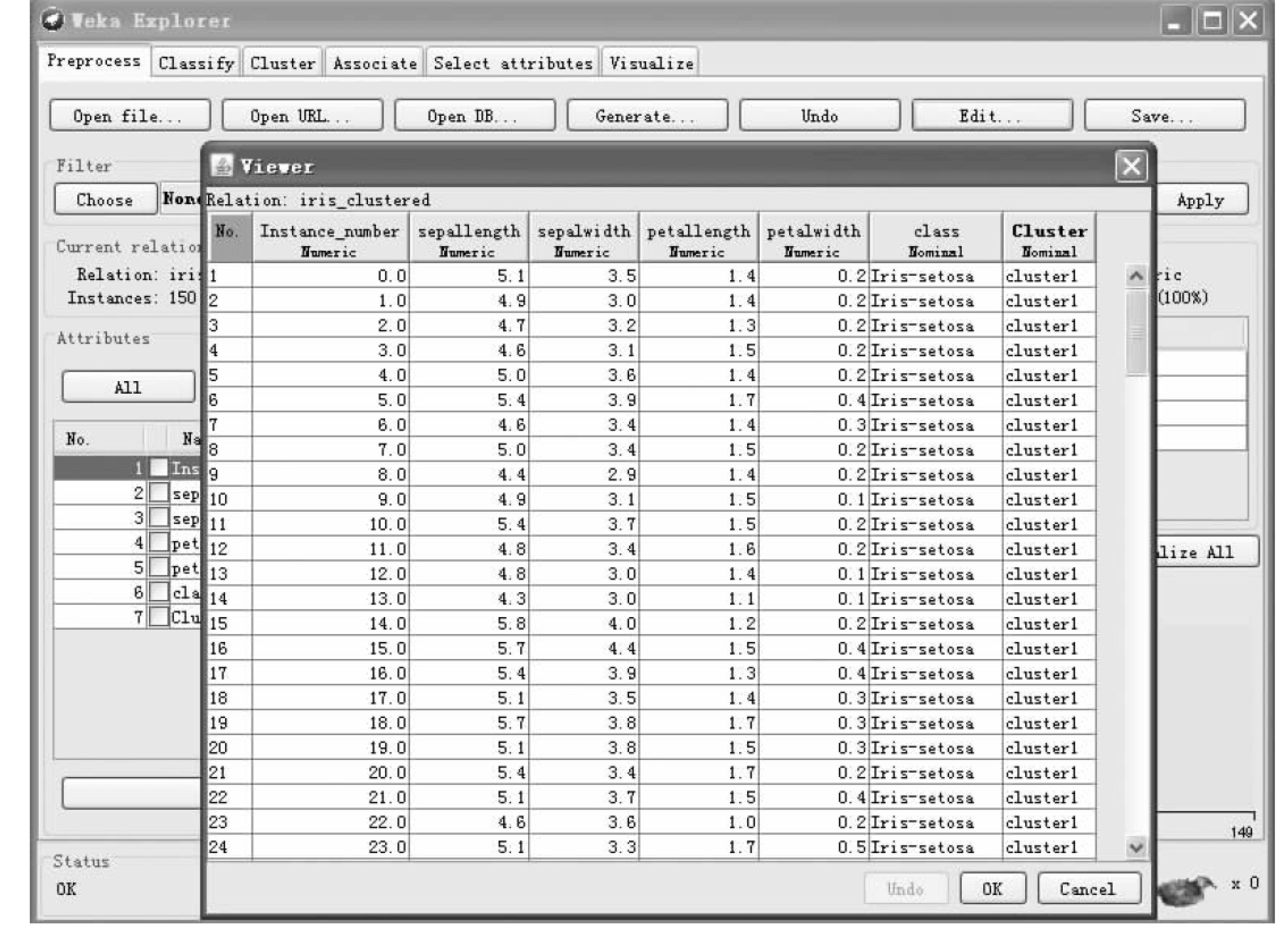


图 8-18　查看每个实例所在簇

8.4　案例小结

聚类是数据挖掘中常见的挖掘任务之一。K-means 是聚类方法中一种简单而有效的方法，被评为数据挖掘中的十大算法之一。本案例利用 Weka 软件，采用 K-means 方法，实现了对橘子花数据的自动聚类，取得了较好的效果。在 Weka 软件中还提供了若干其他聚类方法，读者可在橘子花数据集上检验其他聚类方法的可行性。另外，如何自动确定聚类中簇的数目，在聚类算法研究中至今没有取得满意的答案。

实例 9　基于线性回归方法的汽车油耗预测分析

9.1　任 务 描 述

采集到的 398 辆汽车的数据(取自 1970—1982 年)如表 9-1 所示,每辆汽车给出了 8 个属性值,分别为汽缸(cylinders)、排量(displacement)、马力(horsepower)、重量(weight)、加速度(acceleration)、年份(model)、产地(origin)及油耗(class)。请创建一个能基于汽车的几个特性来推测其油耗(每加仑英里数,MPG)的回归模型。

表 9-1　采集的有关汽车数据

No.	cylinders Nominal	displacement Numeric	horsepower Numeric	weight Numeric	acceleration Numeric	model Nominal	origin Nominal	class Numeric
1	8	307.0	130.0	3504.0	12.0	70	1	18.0
2	8	350.0	165.0	3693.0	11.5	70	1	15.0
3	8	318.0	150.0	3436.0	11.0	70	1	18.0
4	8	304.0	150.0	3433.0	12.0	70	1	16.0
5	8	302.0	140.0	3449.0	10.5	70	1	17.0
6	8	429.0	198.0	4341.0	10.0	70	1	15.0
7	8	454.0	220.0	4354.0	9.0	70	1	14.0
8	8	440.0	215.0	4312.0	8.5	70	1	14.0
9	8	455.0	225.0	4425.0	10.0	70	1	14.0
10	8	390.0	190.0	3850.0	8.5	70	1	15.0
11	8	383.0	170.0	3563.0	10.0	70	1	15.0
12	8	340.0	160.0	3609.0	8.0	70	1	14.0
13	8	400.0	150.0	3761.0	9.5	70	1	15.0
14	8	455.0	225.0	3086.0	10.0	70	1	14.0
15	4	113.0	95.0	2372.0	15.0	70	3	24.0
16	6	198.0	95.0	2833.0	15.5	70	1	22.0
17	6	199.0	97.0	2774.0	15.5	70	1	18.0
18	6	200.0	85.0	2587.0	16.0	70	1	21.0
19	4	97.0	88.0	2130.0	14.5	70	3	27.0
20	4	97.0	46.0	1835.0	20.5	70	2	26.0
21	4	110.0	87.0	2672.0	17.5	70	2	25.0
22	4	107.0	90.0	2430.0	14.5	70	2	24.0
23	4	104.0	95.0	2375.0	17.5	70	2	25.0
24	4	121.0	113.0	2234.0	12.5	70	2	26.0

9.2　技 术 原 理

回归分析是研究变量之间相关关系的一种统计推断法。回归分析,是指在相关分析的基础上,把变量之间的具体变动关系模型化,求出关系方程式,即一个能够反映变量间变化关系的函数关系式,并据此进行估计和推算。通过回归分析,可以将相关变量之间不确定、不规则的数量关系一般化、规范化,从而可以根据自变量的某一个给定值推断出因变量的估计值。根据所涉及变量的多少不同,回归分析可分为一元回归和多元回归。

假设一个随机变量 Y 与 m 个非随机变量 $X_1, X_2, \cdots, X_m$ 之间存在线性相关关系,则它们之间的关系可以用以下线性回归模型来表示

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + e$$

其中,Y 是因变量, $X_i(i=1,2,\cdots,m)$是自变量, $\beta_i(i=0,1,2,\cdots,m)$是模型的参数,称为偏相关系数;$e$ 是随机误差。

回归参数 $\beta_i(i=0,1,2,\cdots,m)$ 的估计方法是最小二乘法。根据样本数据 $(y, x_{1j}, x_{2j}, \cdots, x_{mj})$ 来估计 $\beta_i(i=0,1,2,\cdots,m)$ 时要使得产生残差的平方和为

$$Q = \sum(y_j - \hat{y}_j)^2 = \sum[y_j - (\beta_0 + \beta_1 x_{1j} + \cdots + \beta_m x_{mj})]^2 \tag{9-1}$$

取极小值。为此，对 Q 分别求 $\beta_i(i=0,1,2,\cdots,m)$ 的偏导数，并令其等于零，由此可以得到 $m+1$ 个方程。

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{j=1}^{n}[y_j - (\beta_0 + \beta_1 x_{1j} + \cdots + \beta_m x_{mj})] = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{j=1}^{n}[y_j - (\beta_0 + \beta_1 x_{1j} + \cdots + \beta_m x_{mj})]x_{1j} = 0 \\ \cdots \\ \dfrac{\partial Q}{\partial \beta_m} = -2\sum_{j=1}^{n}[y_j - (\beta_0 + \beta_1 x_{1j} + \cdots + \beta_m x_{mj})]x_{mj} = 0 \end{cases} \tag{9-2}$$

整理后可得方程组

$$\begin{cases} n\beta_0 + \sum_{j=1}^{n} x_{1j}\beta_1 + \cdots + \sum_{j=1}^{n} x_{mj} = \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} x_{1j}\beta_0 + \sum_{j=1}^{n} x_{1j}^2\beta_1 + \cdots + \sum_{j=1}^{n} x_{1j}x_{m}\beta_m = \sum_{j=1}^{n} x_{1j}y_j \\ \cdots \\ \sum_{j=1}^{n} x_{mj}\beta_0 + \sum_{j=1}^{n} x_{mj}x_{1j}\beta_1 + \cdots + \sum_{j=1}^{n} x_{mj}^2\beta_m = \sum_{j=1}^{n} x_{mj}y_j \end{cases} \tag{9-3}$$

对于自变量 $X_1, X_2, \cdots, X_m$ 和因变量 Y 共有 n 组观察数据。x_{ik} 表示自变量 X_i 的第 k 次观察值，y_i 表示因变量 Y 的第 i 次观察值。令

$$l_{ij} = \sum_{k=1}^{n}(x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \quad (i,j = 1,2,\cdots,m)$$

$$l_{i0} = \sum_{k=1}^{n}(x_{ik} - \overline{x_i})(y_k - \overline{y}) \quad (i = 1,2,\cdots,m) \quad l_{00} = \sum_{k=1}^{n}(y_k - \overline{y})^2$$

则回归系数 $\beta_i(i=0,1,2,\cdots,m)$ 可以由方程组求出

$$\begin{cases} l_{11}\beta_1 + l_{12}\beta_2 + \cdots + l_{1m} = l_{10} \\ l_{21}\beta_1 + l_{22}\beta_2 + \cdots + l_{2m} = l_{20} \\ \cdots \\ l_{m1}\beta_1 + l_{m2}\beta_2 + \cdots + l_{mm} = l_{m0} \end{cases} \tag{9-4}$$

常数项 $\beta_0 = Y - \sum\beta_i \cdot \overline{x_i}$。

9.3 具 体 实 现

(1) 依次单击“开始”→“所有程序”→Weka 3.6.5→Weka 3.6，如图 9-1 所示。

(2) 单击 Explorer 按钮，如图 9-2 所示。

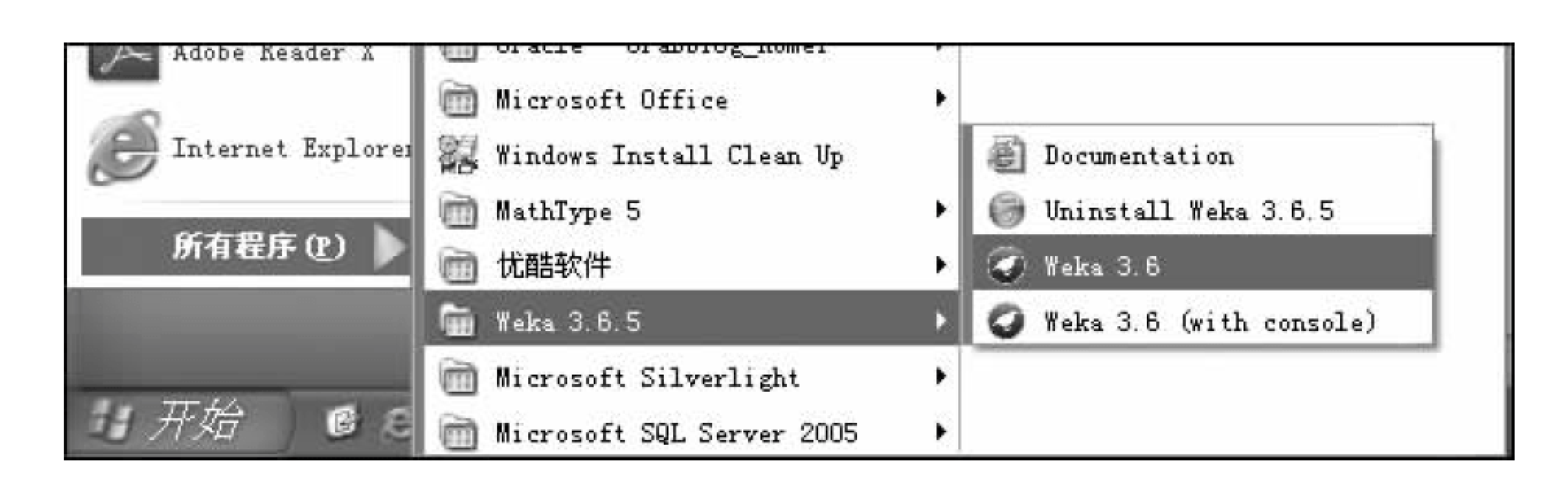


图 9-1　打开 Weka 软件

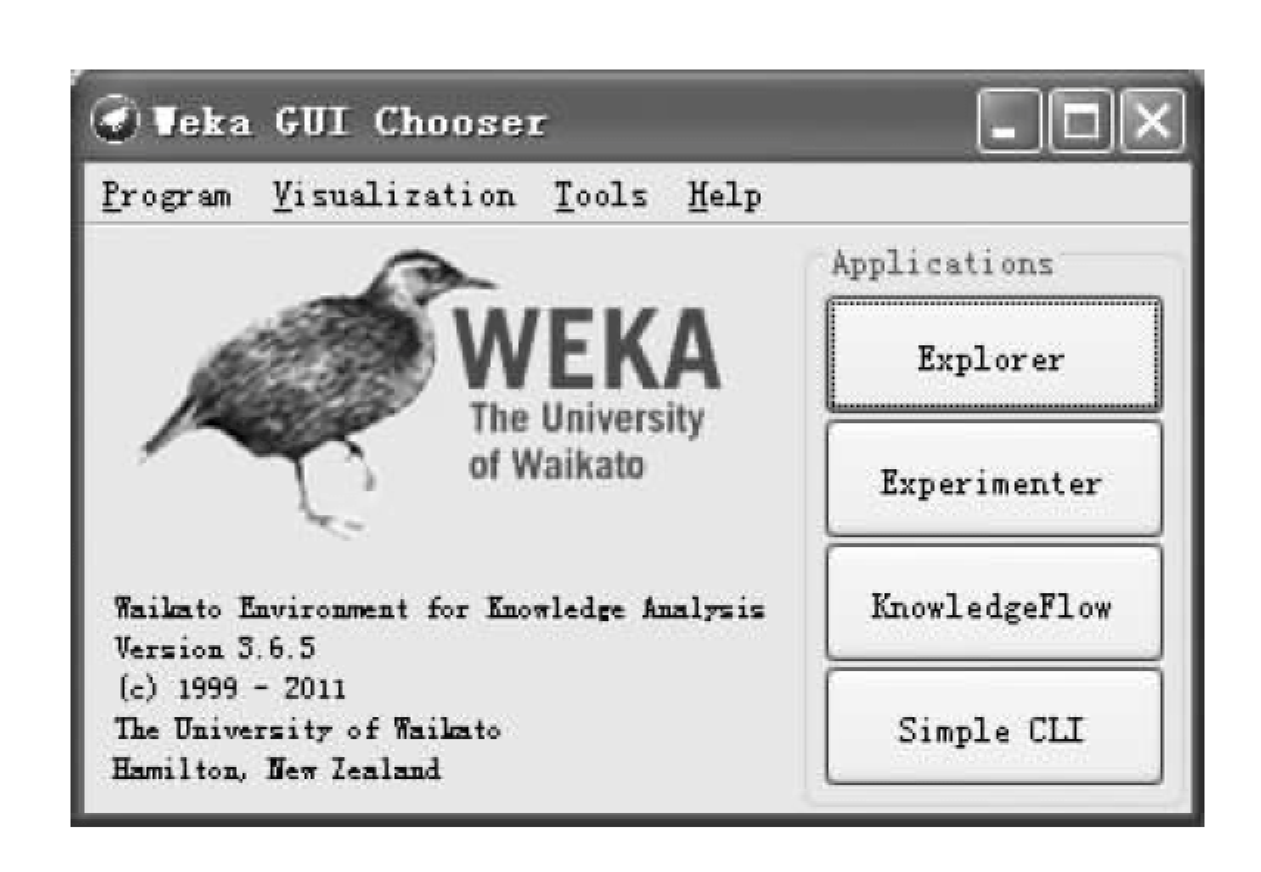


图 9-2　打开 Explorer 应用

(3) 单击 Open file 按钮,选择要打开的文件 autoMpg. arff,并单击“打开”按钮,如图 9-3 所示。

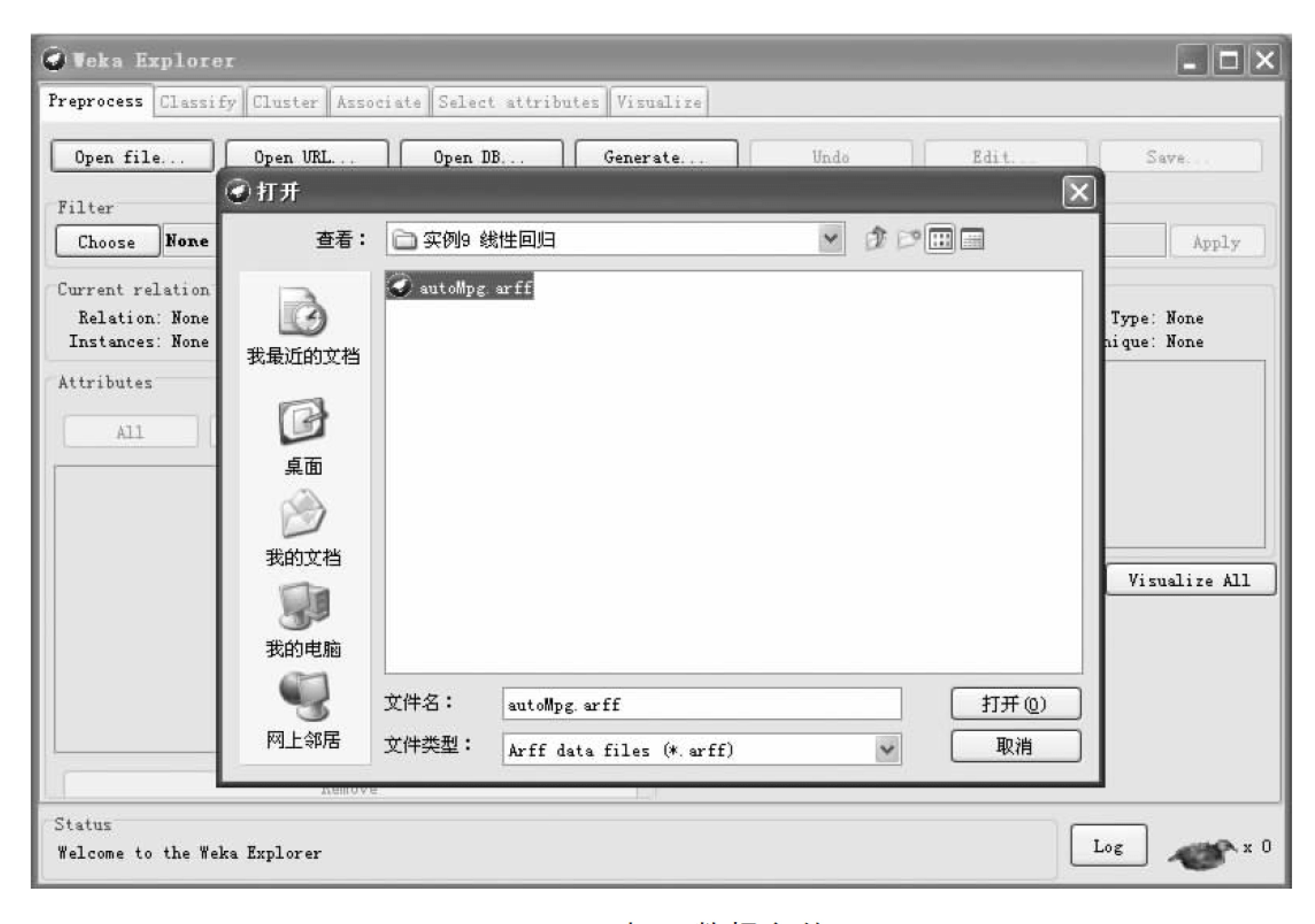


图 9-3　打开数据文件

(4) 在如图 9-4 所示的界面中,可以知道 autoMpg 数据集中共有 398 个实例,每个实例有 8 个属性。选中某个属性,可以查看 398 个实例关于这个属性的属性值的最小值、最大值、均值和标准差等信息。

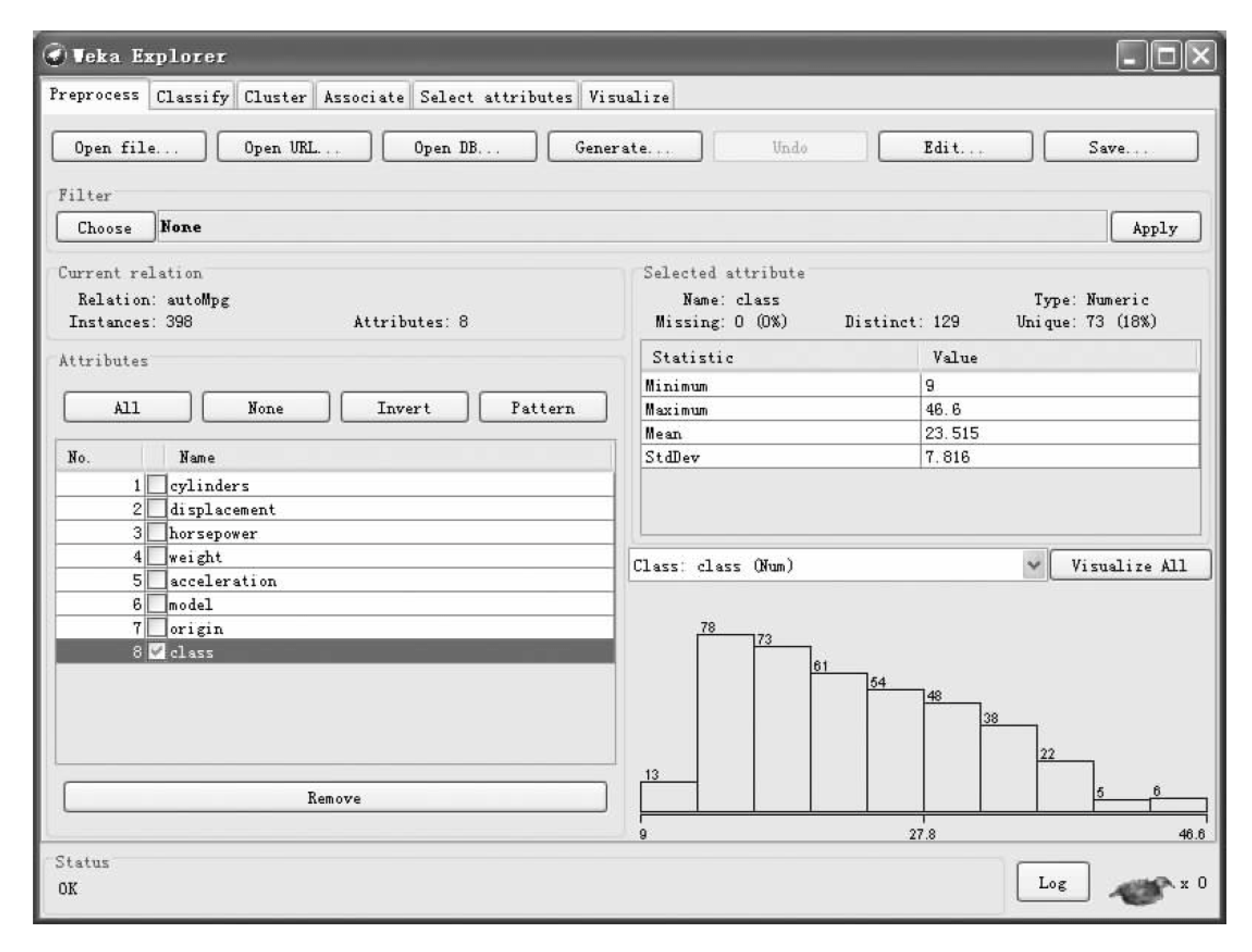


图 9-4　查看数据特征

（5）选择 Classify 标签，并单击 Choose 按钮，如图 9-5 所示。

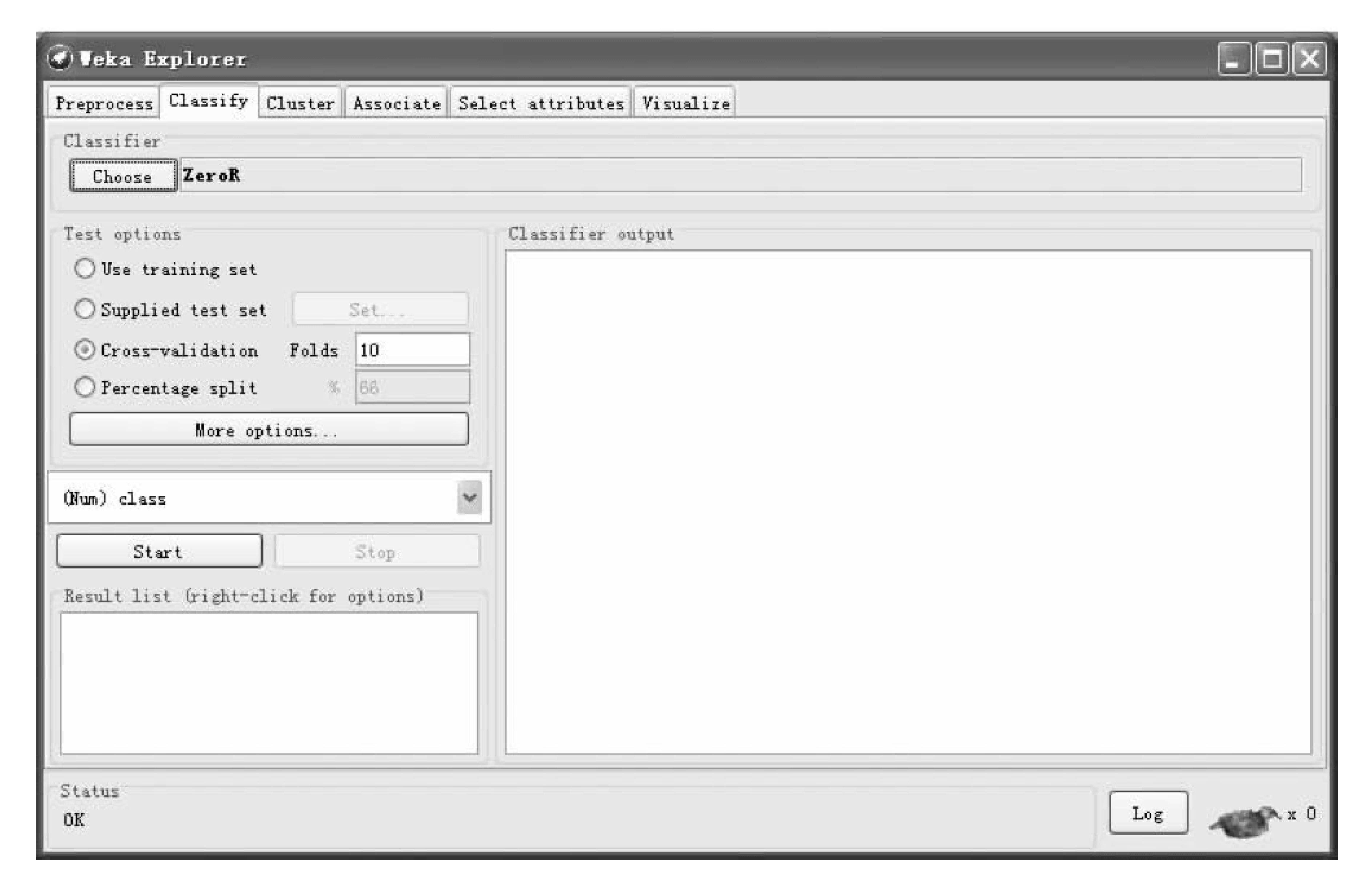


图 9-5　选择 Classify 标签

（6）选择 LinearRegression 方法，并单击 Close 按钮，如图 9-6 所示。

（7）单击 Choose 按钮后的 LinearRegression 方法，弹出参数设置框，这里选择默认参数，并单击 OK 按钮，如图 9-7 所示。

（8）在 Test options 选项中选择 Use training set 单选按钮，并将“油耗(class)”设为因

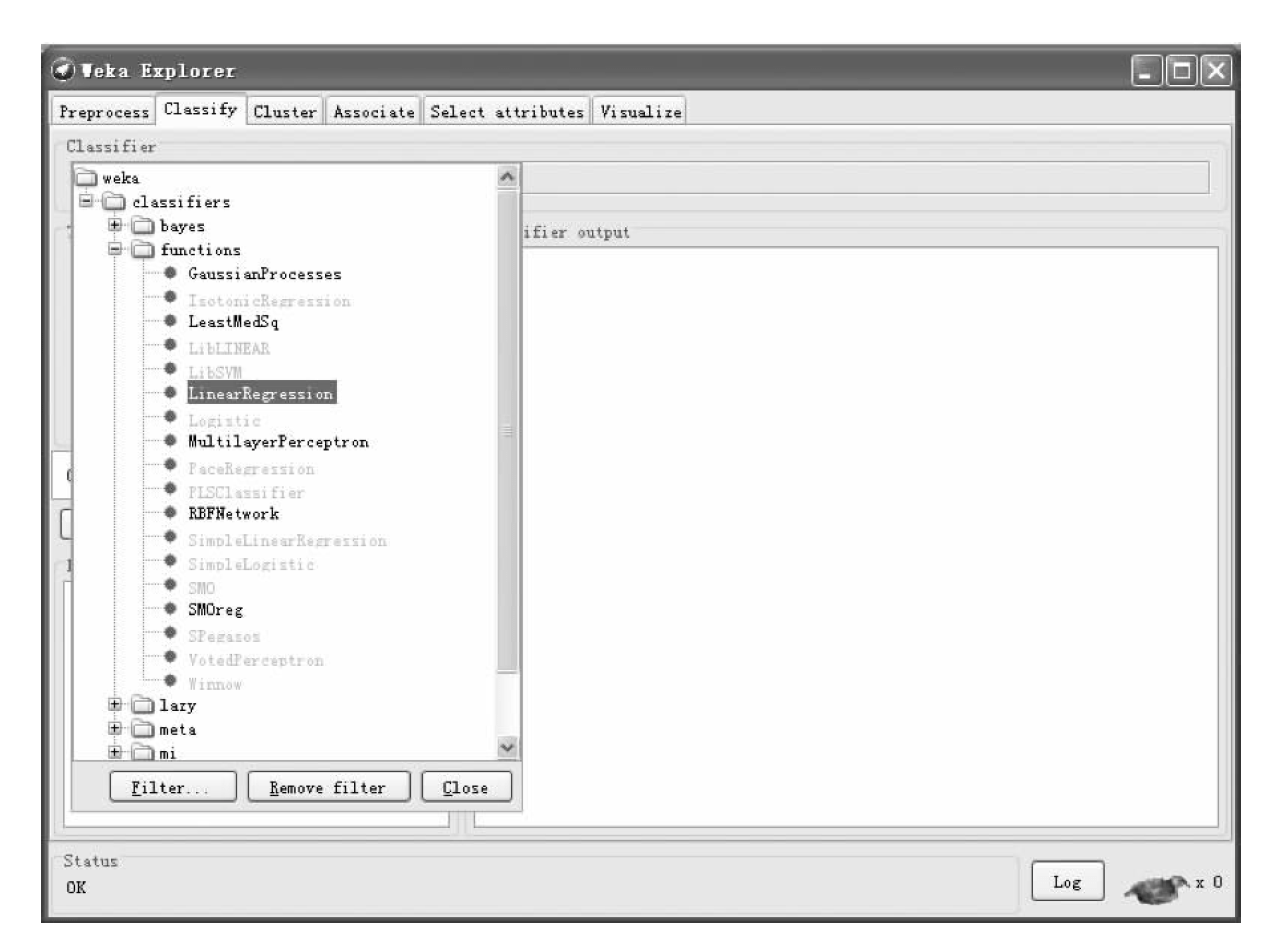


图 9-6　选择 LinearRegression 方法

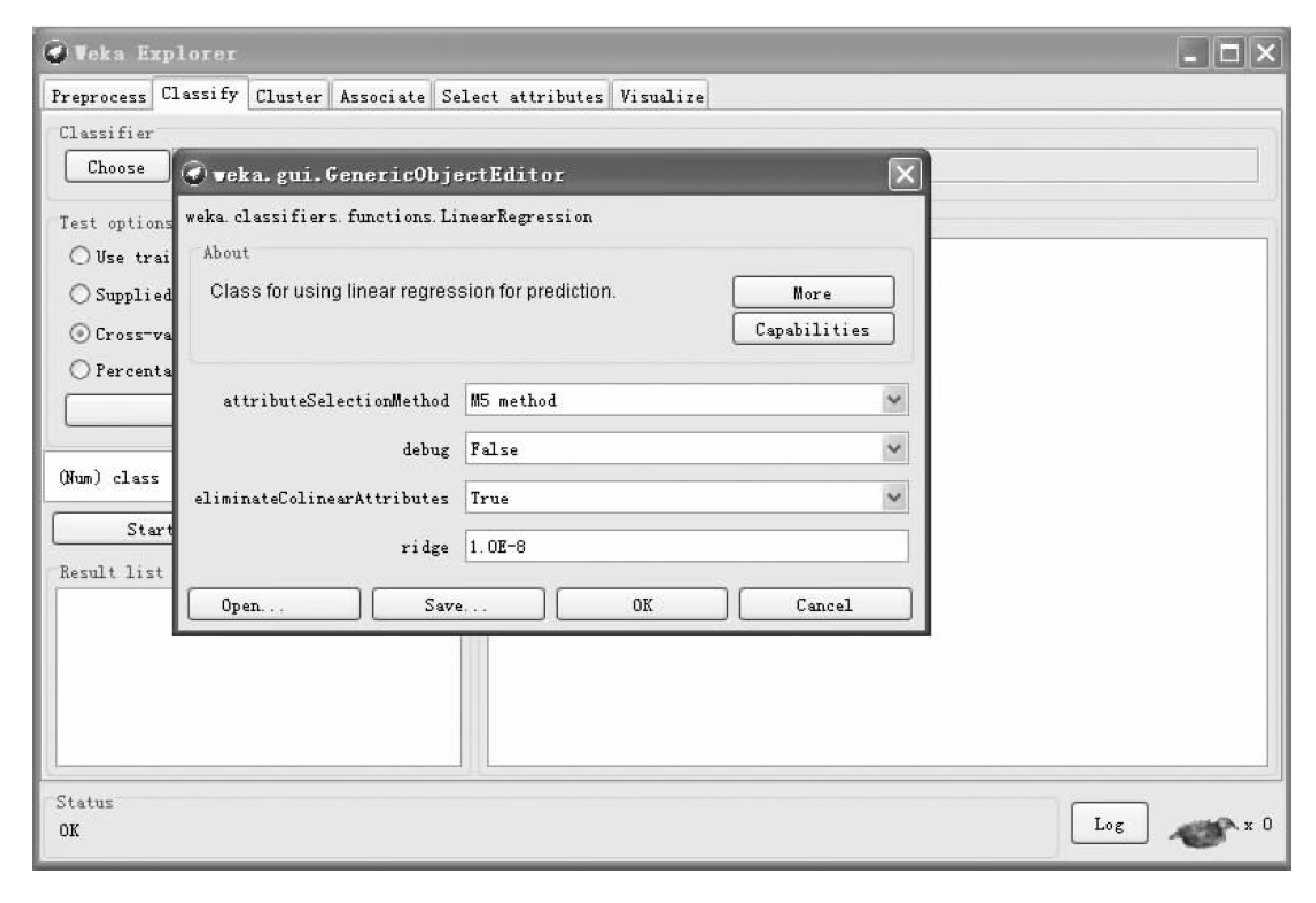


图 9-7　进行参数设置

变量，如图 9-8 所示。

（9）单击 Start 按钮，Weka 对 autoMpg 数据集执行线性回归算法，在算法的执行结果中给出了建立的线性回归模型，如图 9-9 所示。这个回归模型的首行，−2.2744 * cylinders=6，3，5，4 表示，如果汽车有 6 个缸、3 个缸、5 个缸或 4 个缸，就会在此列中放上一个 1，如果为其他缸数，就会放上一个 0。

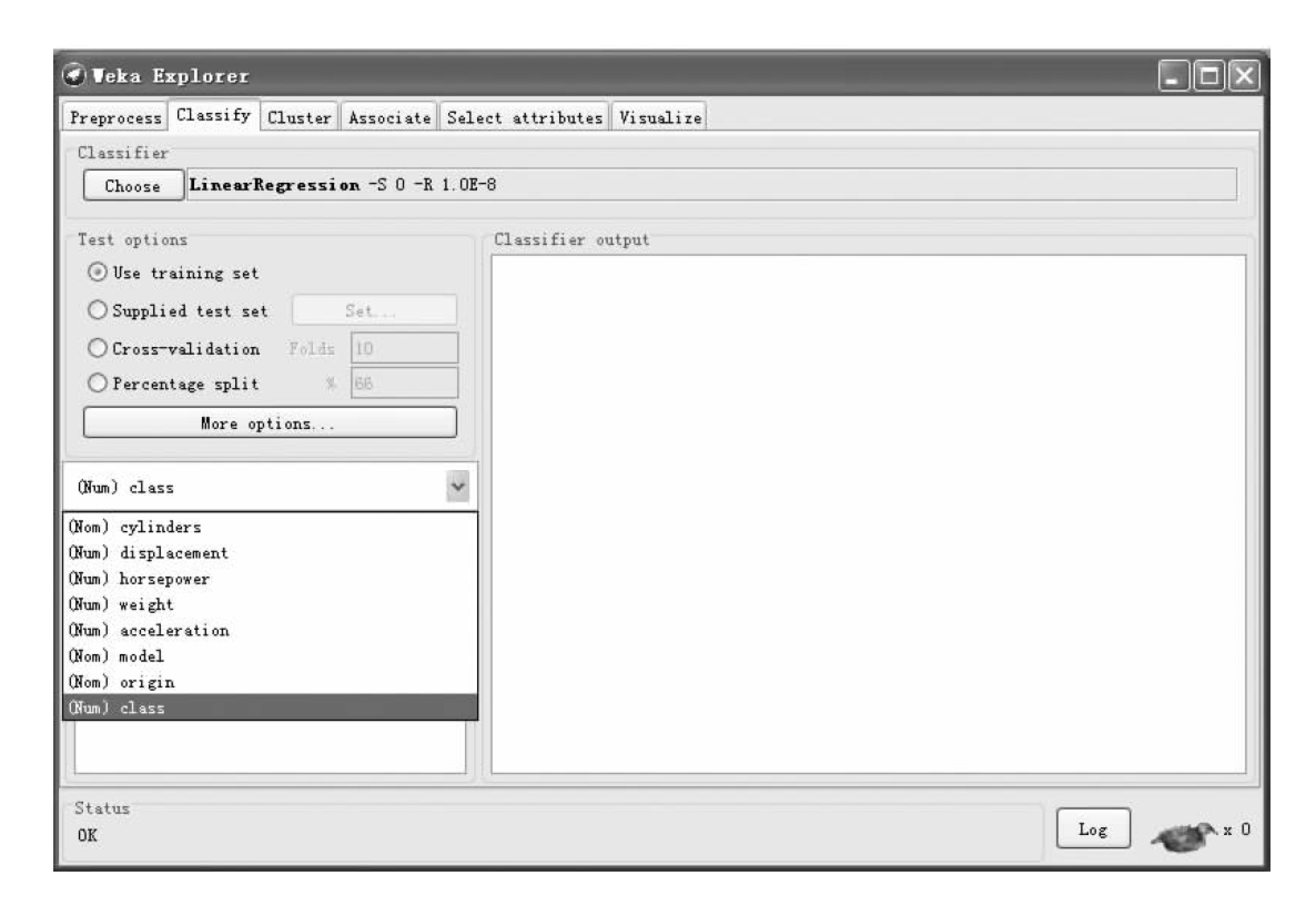


图 9-8　设置因变量

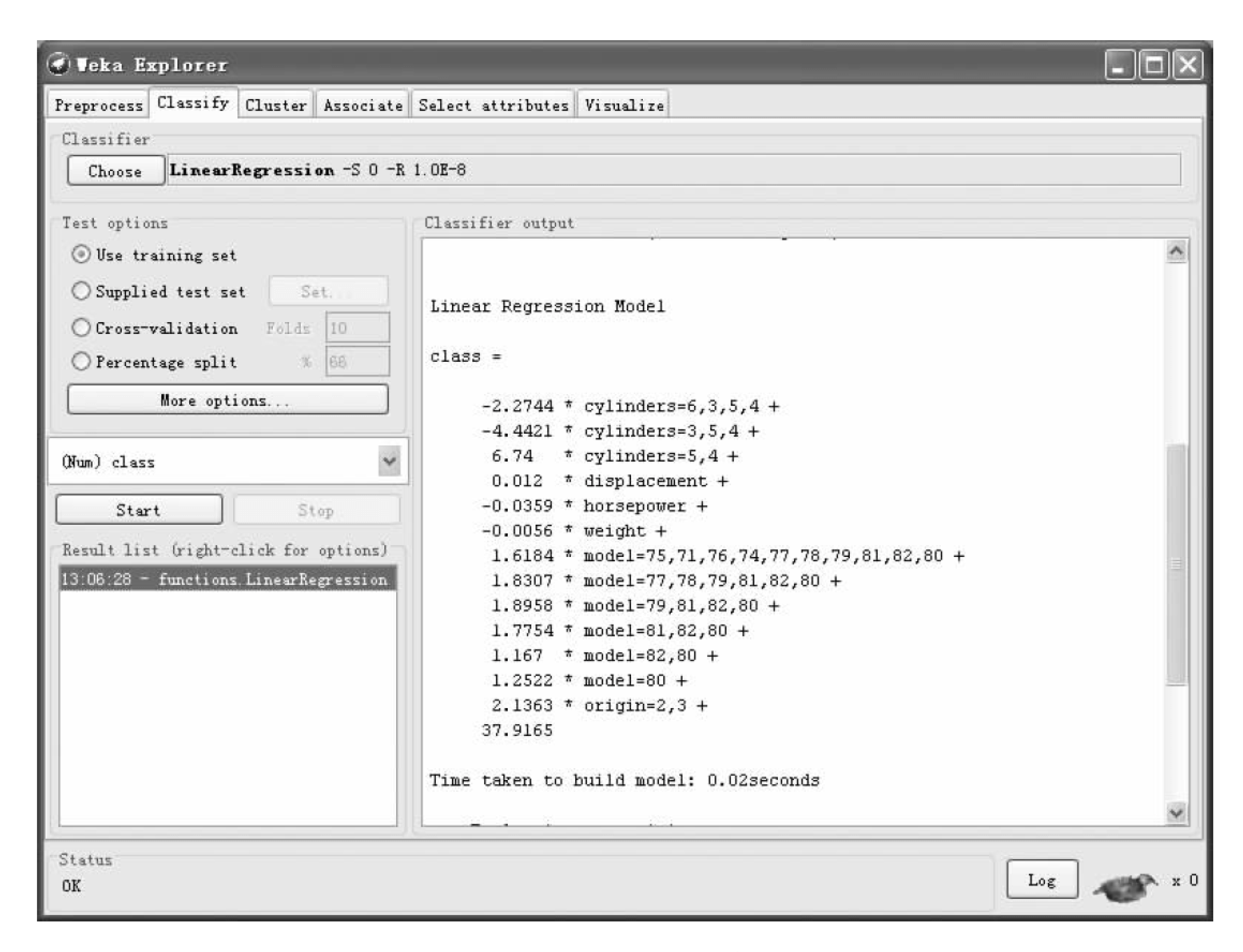


图 9-9　执行回归算法

（10）为了观察可视化的分类结果，在左下方 Result list 列出的结果上右击，在弹出菜单中选择 Visualize classifier errors 项，如图 9-10 所示。

（11）在弹出的可视化结果对话框中，可以查看实际油耗值和预测油耗值之间的对应关系，如图 9-11 所示。

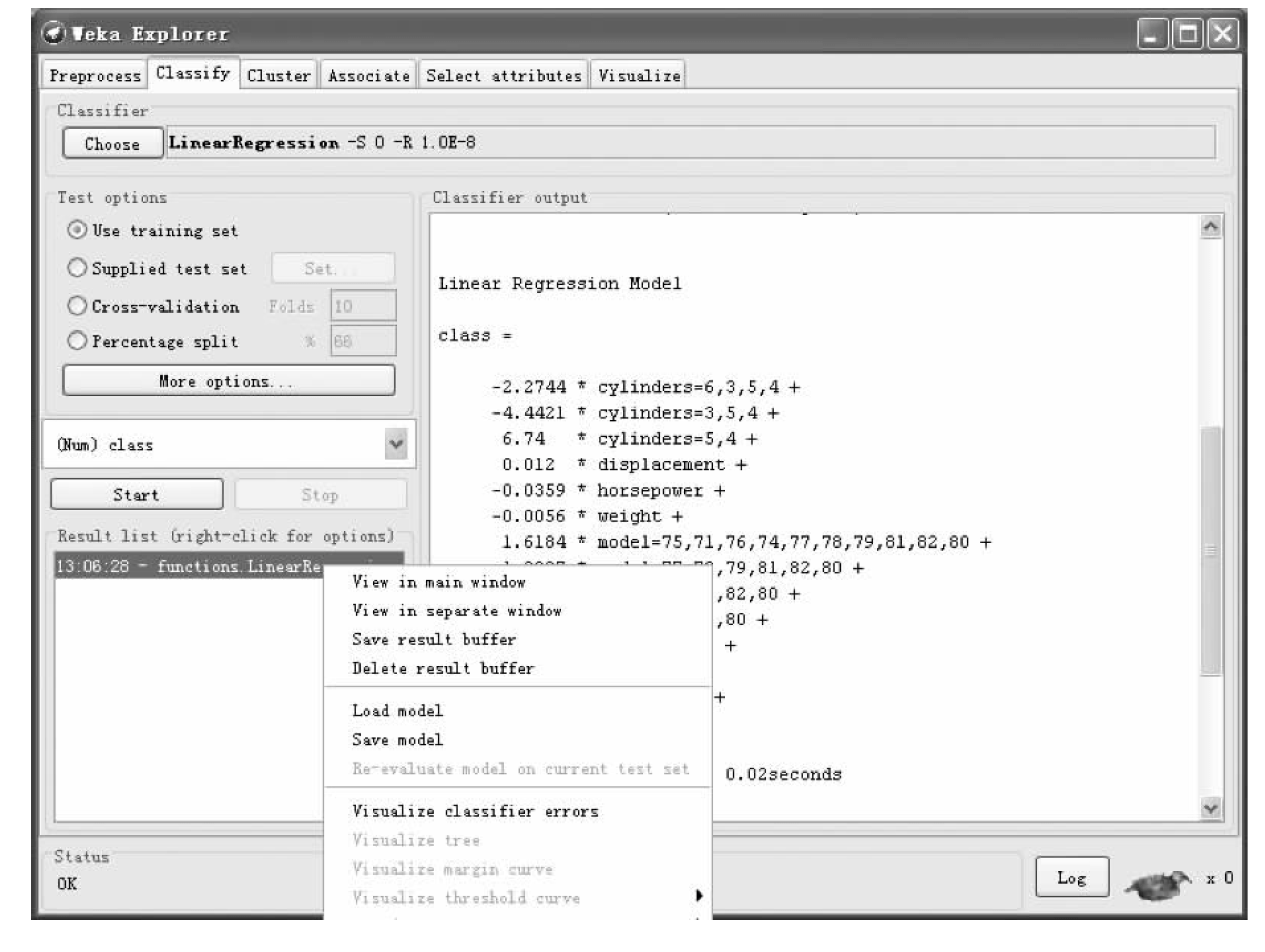


图 9-10 选择 Visualize classifier errors 选项

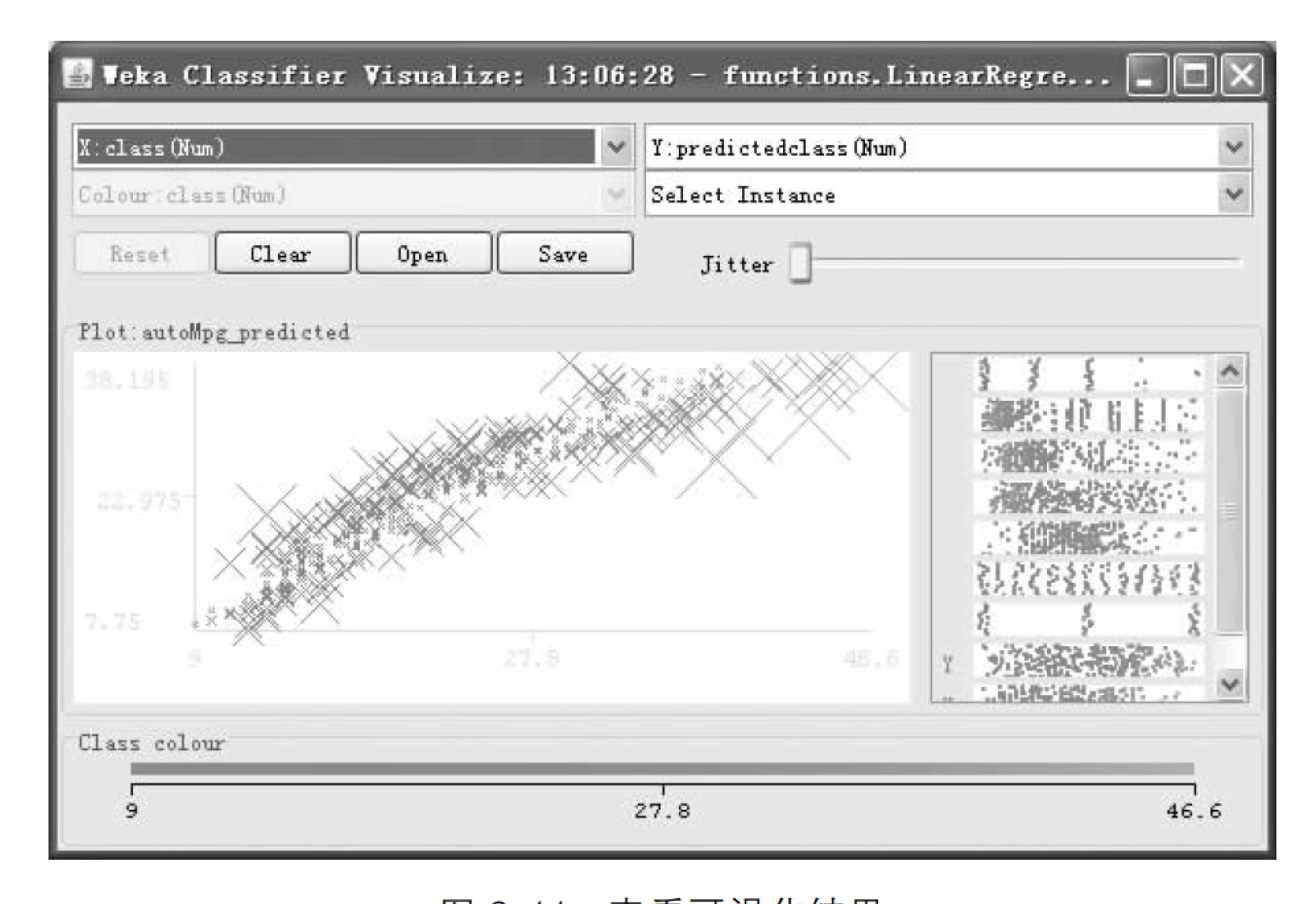


图 9-11 查看可视化结果

（12）在可视化结果对话框中，单击 Save 按钮，将结果保存在 autoresult.arff 文件中，如图 9-12 所示。

（13）在 Weka Explorer 界面中打开 autoresult.arff 文件，如图 9-13 所示。在此界面中，可以查看每个对象的实际油耗值和预测油耗值。例如，编号为 10 的汽车，实际油耗值为 15.0，预测油耗值为 14.356 11。

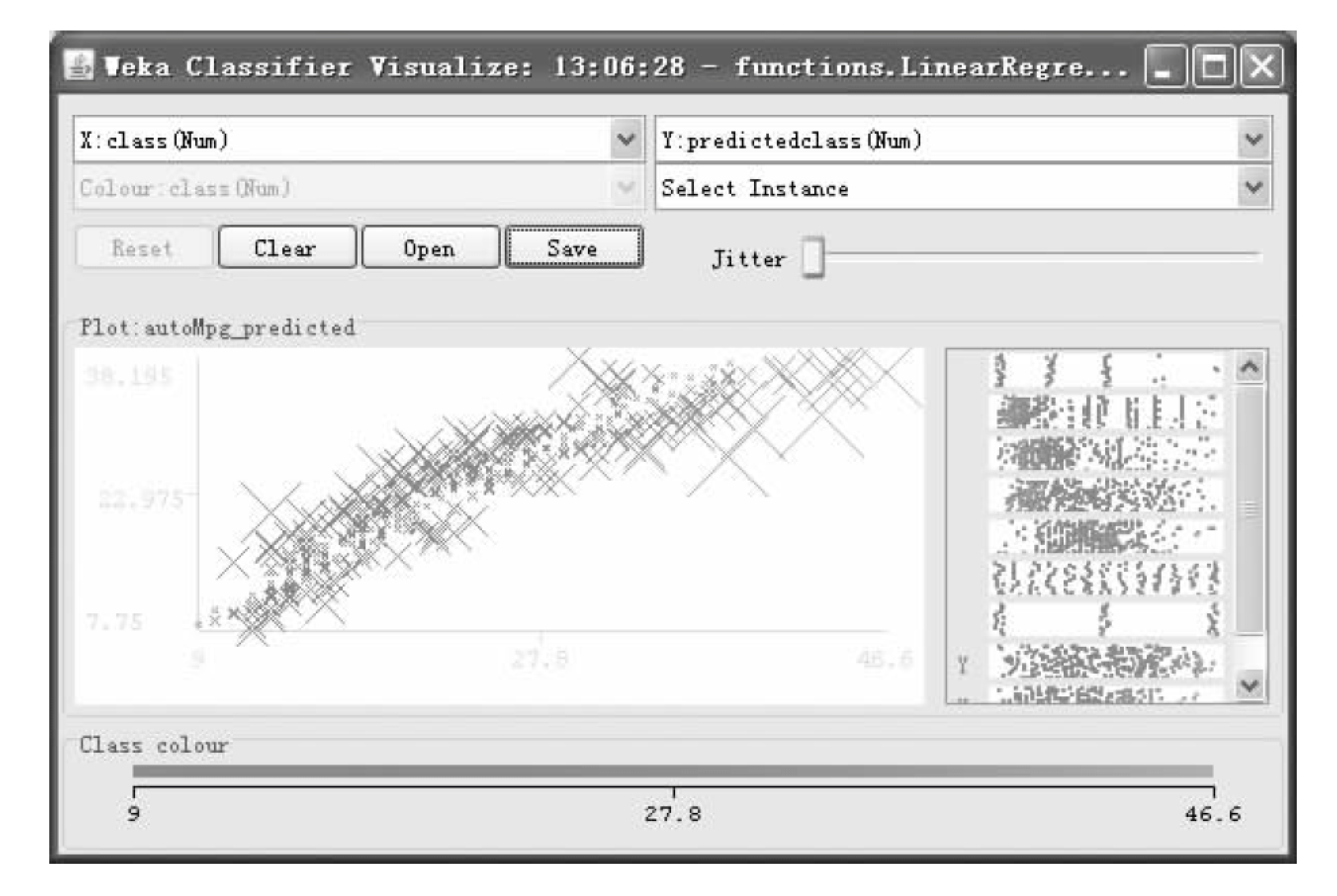
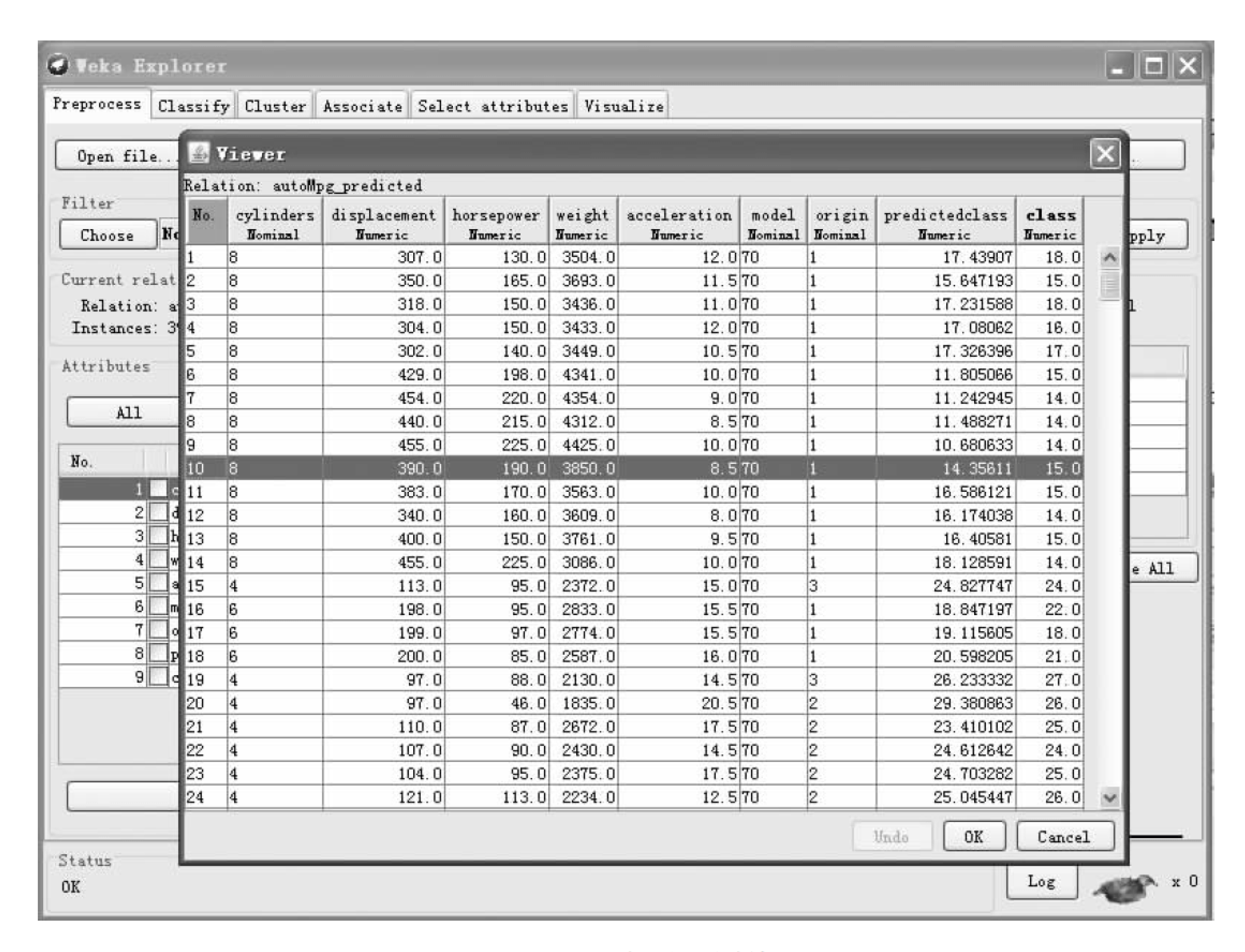


图 9-12　保存可视化结果



No.	cylinders Nominal	displacement Numeric	horsepower Numeric	weight Numeric	acceleration Numeric	model Nominal	origin Nominal	predictedclass Numeric	class Numeric
1	8	307.0	130.0	3504.0	12.0	70	1	17.43907	18.0
2	8	350.0	165.0	3693.0	11.5	70	1	15.647193	15.0
3	8	318.0	150.0	3436.0	11.0	70	1	17.231588	18.0
4	8	304.0	150.0	3433.0	12.0	70	1	17.08062	16.0
5	8	302.0	140.0	3449.0	10.5	70	1	17.326396	17.0
6	8	429.0	198.0	4341.0	10.0	70	1	11.805066	15.0
7	8	454.0	220.0	4354.0	9.0	70	1	11.242945	14.0
8	8	440.0	215.0	4312.0	8.5	70	1	11.488271	14.0
9	8	455.0	225.0	4425.0	10.0	70	1	10.680633	14.0
10	8	390.0	190.0	3850.0	8.5	70	1	14.35811	15.0
11	8	383.0	170.0	3563.0	10.0	70	1	16.586121	15.0
12	8	340.0	160.0	3609.0	8.0	70	1	16.174038	14.0
13	8	400.0	150.0	3761.0	9.5	70	1	16.40581	15.0
14	8	455.0	225.0	3086.0	10.0	70	1	18.128591	14.0
15	4	113.0	95.0	2372.0	15.0	70	3	24.827747	24.0
16	6	198.0	95.0	2833.0	15.5	70	1	18.847197	22.0
17	6	199.0	97.0	2774.0	15.5	70	1	19.115605	18.0
18	6	200.0	85.0	2587.0	16.0	70	1	20.598205	21.0
19	4	97.0	88.0	2130.0	14.5	70	3	26.233332	27.0
20	4	97.0	46.0	1835.0	20.5	70	2	29.380863	26.0
21	4	110.0	87.0	2672.0	17.5	70	2	23.410102	25.0
22	4	107.0	90.0	2430.0	14.5	70	2	24.612642	24.0
23	4	104.0	95.0	2375.0	17.5	70	2	24.703282	25.0
24	4	121.0	113.0	2234.0	12.5	70	2	25.045447	26.0

图 9-13　查看预测值

9.4　案例小结

回归分析是研究变量之间相关关系的一种统计推断法。根据所涉及变量的多少不同，回归分析可分为一元回归和多元回归。本案例利用 Weka 软件，采用线性回归方法，实现了对汽车油耗的预测，取得了较好的效果。Weka 软件中提供的线性回归算法，不仅能够处理数值型属性，还能处理名词性属性（如本案例中的汽缸数），这是传统线性回归方法的一个拓展。

实例 10　基于决策树方法的中文文本自动分类分析

10.1　任 务 描 述

在某个门户网站上搜集到了 120 篇文本，其中 1～40 篇属于军事类；41～80 篇属于电影类，81～120 篇属于篮球类。文本集存储在文本文件“军事-电影-篮球.txt”中。在每类中各选取 30 篇文本，共 90 篇文本作为训练集，利用文本分类技术建立分类器，预测剩余 30 篇文本所属的类别。并利用 30 篇测试文本的已知类别，评价所建立文本分类器在正确率方面的性能。

10.2　技 术 原 理

10.2.1　文本挖掘的概念

在现实世界，面对的数据大都是文本数据，由各种数据源（如新闻文章、研究论文、书籍、数字图书馆、电子邮件和 Web 页面）的大量文本组成。文本数据不同于传统数据库中的数据，它具有自己的特点：

半结构化：文本数据既不是完全无结构的也不是完全结构化的。例如，文本可能包含结构字段，如标题、作者、出版日期、长度、分类等，也可能包含大量的非结构化的数据，如摘要和内容。

高维：文本向量的维数一般都可以高达上万维，一般的数据挖掘、数据检索的方法由于计算量过大或代价高昂而不具有可行性（比如多元统计分析中的主因素分析）。

高数据量：一般的文本库中都会存在最少数千个文本样本，对这些文本进行预处理、编码、挖掘等处理的工作量是非常庞大的，因而手工方法一般是不可行的。

语义性：文本数据中存在着一词多义、多词一义，在时间和空间上的上下文相关等情况。

随着信息技术的发展，文本数据的数量急剧增长，所以对文本进行数据挖掘成为了数据挖掘的一个发展方向。文本挖掘以文本型信息源作为分析的对象，利用定量计算和定性分析的方法，从中寻找到信息结构、模型、模式等各种隐含的新颖知识。

文本挖掘过程一般包括文本分词、文本特征表示、词频矩阵降维、文本相似度计算、文本知识获取等。在经过对文本数据进行一系列的预处理以后，传统的数据挖掘方法同样可以应用于文本数据挖掘。

10.2.2　文本分词技术

分词是中文信息处理从字符处理水平向语义处理水平迈进的关键。汉语文本不像西文那样，词与词之间有空格间隔，同时由于汉语的构词方式、不同分词方式表达不同意义等特

点，使得中文处理必须有分词这道工序。

汉语分词的难点主要表现在两个方面，即歧义切分和未登录词的切分。

- 歧义切分：汉语字与字之间组词灵活，给分词带来了很大的困难。从上下文关系的角度看，其中只有一种切分结果是正确的。
- 未登录词切分：未登录词主要是指分词系统的词典中未收录的词。不断出现的新词属于另外一类未登录词，反映在自然语言上就是大量的新词不断涌现。

分词技术，大致可以分为 5 类：词典分词法、切分标记分词法、基于统计的分词方法、基于语言规则的分词方法和智能分词方法。

10.2.3 文本特征表示

文本特征指的是关于文本的元数据，分为两种：描述性特征，如文本的名称、日期、大小、类型等；语义性特征，如文本的作者、机构、标题、内容等。描述性特征易于获得，而语义性特征则较难得到。

向量空间模型是近年来应用较多且效果较好的表示文本特征的方法。在该模型中，文本空间被看作是由一组正交词条向量所张成的空间，每一个词条称为一个特征项，每一个文本 $\boldsymbol{d}$ 则表示为空间内的一个向量，一般表示为

$$V(\boldsymbol{d}) = (w(t_1), w(t_2), w(t_3), \cdots, w(t_n)) \tag{10-1}$$

其中，t_i 为张成文本空间的词条；n 为文本空间的维数；$w(t_i)$是函数，其基本功能是计算词条 t_i 在文本向量中的权重；$w(t_i)$一般被定义为 t_i 在文本 d 中出现频率 $tf_i(d)$的函数，即 $w(t_i)=\psi(tf_i(d))$，常用的 ψ 如下：

(1) 布尔函数：

$$\psi = \begin{cases} 1 & tf_i(d) > 0 \\ 0 & tf_i(d) = 0 \end{cases} \tag{10-2}$$

(2) 平方根函数：

$$\psi = \sqrt{tf_i(d)} \tag{10-3}$$

(3) 对数函数：

$$\psi = \log(tf_i(d) + 1) \tag{10-4}$$

(4) TFIDF 函数：

$$\psi = tf_i(d) \times \log\left(\frac{N}{n_i}\right) \tag{10-5}$$

其中 N 为所有文本的数目，n_i 为含有词条 t_i 的文本数目。

10.3 具体实现

(1) 中文分词技术相对比较复杂，这里使用现成的中文分词软件进行分词。在浏览器中输入网址 http://www.hylanda.com/，进入天津海量信息技术有限公司的体验中心，并单击海量中文智能分词选项，进入中文智能分词体验界面，如图 10-1 所示。

(2) 依次将每篇文本粘贴到“输入原文”框，并单击“显示分词结果”按钮，则分词后的结果显示在“分词结果”中，如图 10-2 所示。



图 10-1 进入智能分词体验界面

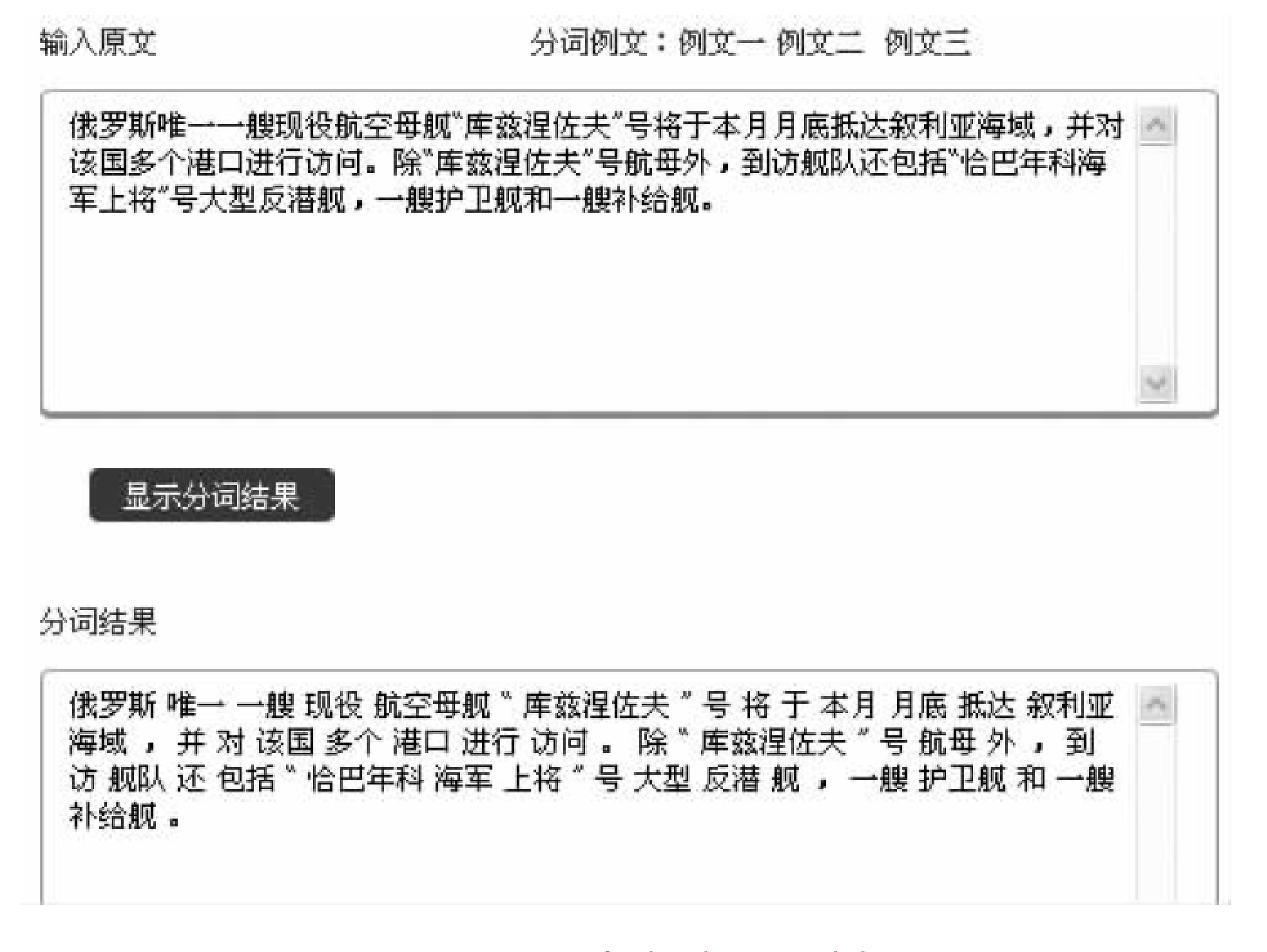


图 10-2 中文分词示例

（3）进行预过滤，以避免非常短或非常长的关键词以及不是单词的词语；再进行高、低通过滤，过滤那些很不常用和诸如辅助动词那些出现频率很高的常用词。以上过程读者可以自己编程实现。分词并进行过滤后的 120 篇文本存储在文件“军事-电影-篮球分词.xlsx”中。

（4）经过统计在 120 篇文本中共出现了 522 个词语。然后统计每个词语在每个文本中出现的次数，形成 120×522 的词频矩阵。将词频矩阵保存在文件“词频矩阵.xlsx”中。

（5）利用公式(10-5)计算每个词语在每篇文本中的权重，形成 120×522 的文本特征向量矩阵。将训练文本集和测试文本集的文本特征向量矩阵分别存储在“文本特征向量矩阵90.csv”和“文本特征向量矩阵 30.csv”中。

（6）分别在文件“文本特征向量矩阵 90.csv”和“文本特征向量矩阵 30.csv”中的最后加入一列，用来表示文本的类别。其中 x 代表此文本属于军事类别，y 代表此文本属于电影类

别，z 代表此文本属于篮球类别。

（7）选择“开始”→“所有程序”→Weka 3.6.5→Weka 3.6 命令，如图 10-3 所示。

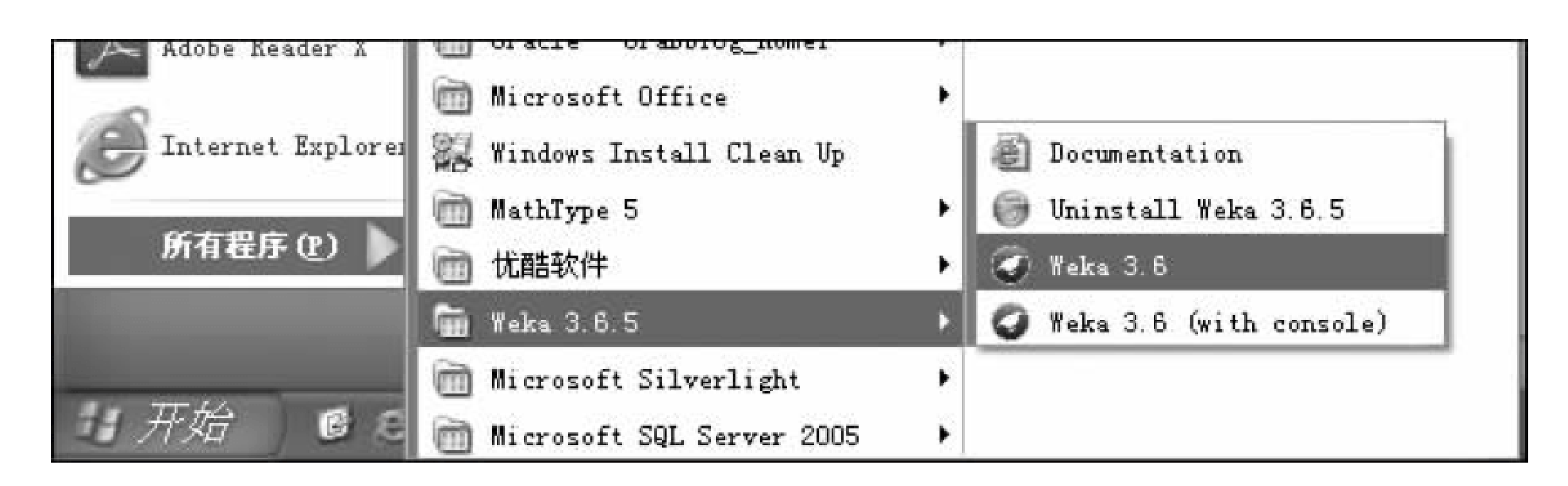


图 10-3 打开 Weka 软件

（8）单击 Explorer 按钮，如图 10-4 所示。

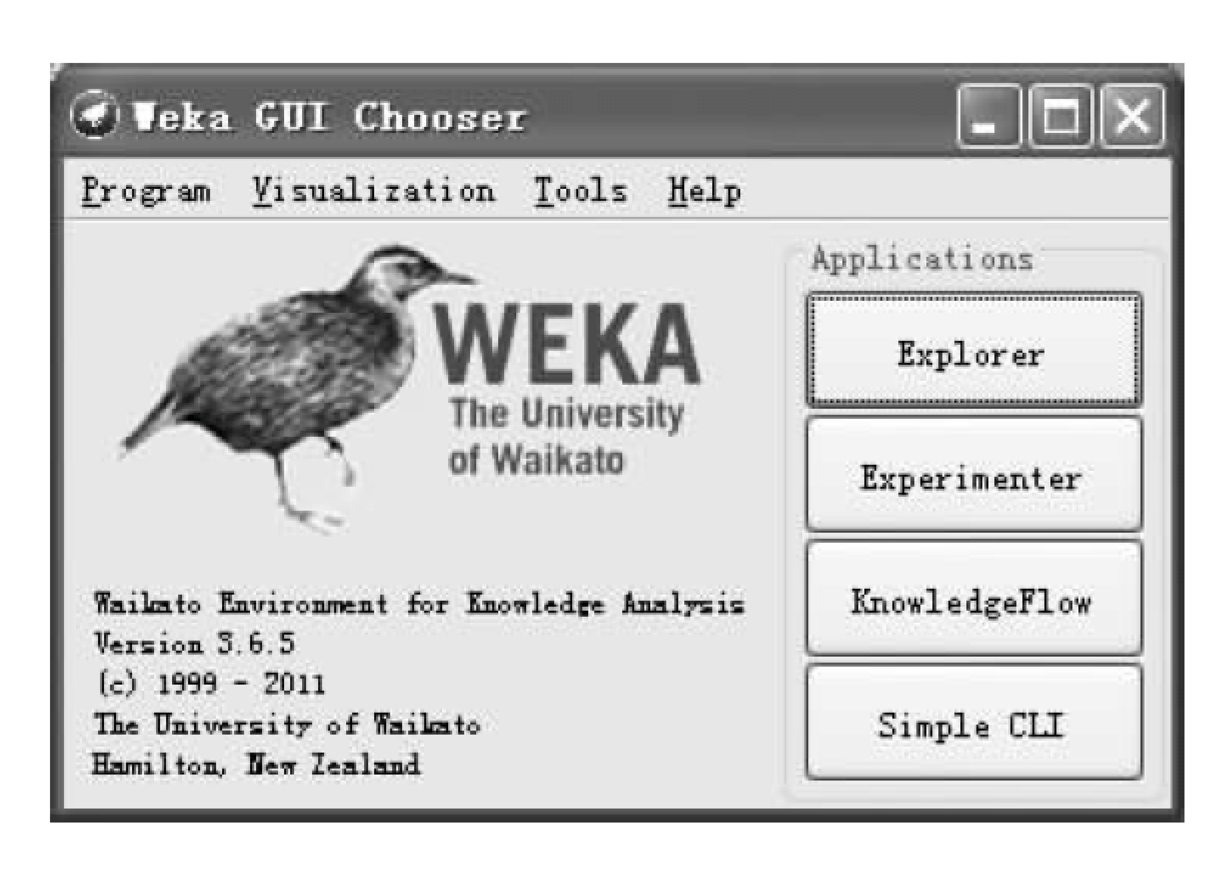


图 10-4 打开 Explorer 应用

（9）单击 Open file 按钮，选择要打开的文件“文本特征向量矩阵 90.csv”，并单击“打开”按钮，如图 10-5 所示。



图 10-5 打开训练数据文件

（10）在如图 10-6 所示的界面中，可以知道此数据集中共有 90 个实例，每个实例有 522 个属性。选中某个属性，可以查看 90 个实例关于此属性的属性值的最小值、最大值、均值和标准差等信息。

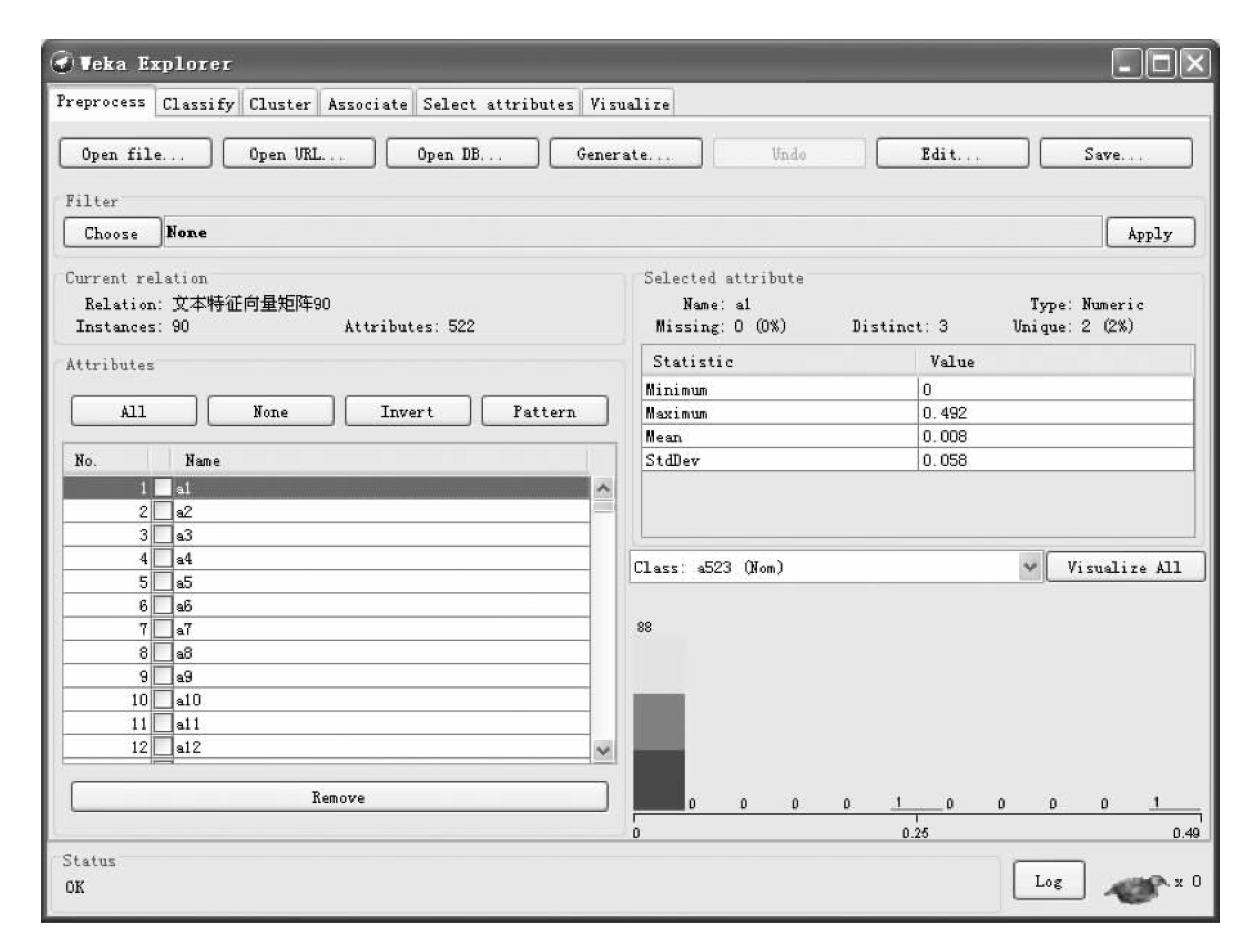


图 10-6　查看数据特征

（11）单击 Classify 标签，并单击 Choose 按钮。如图 10-7 所示，选择 J48 分类器，并单击 Close 按钮。

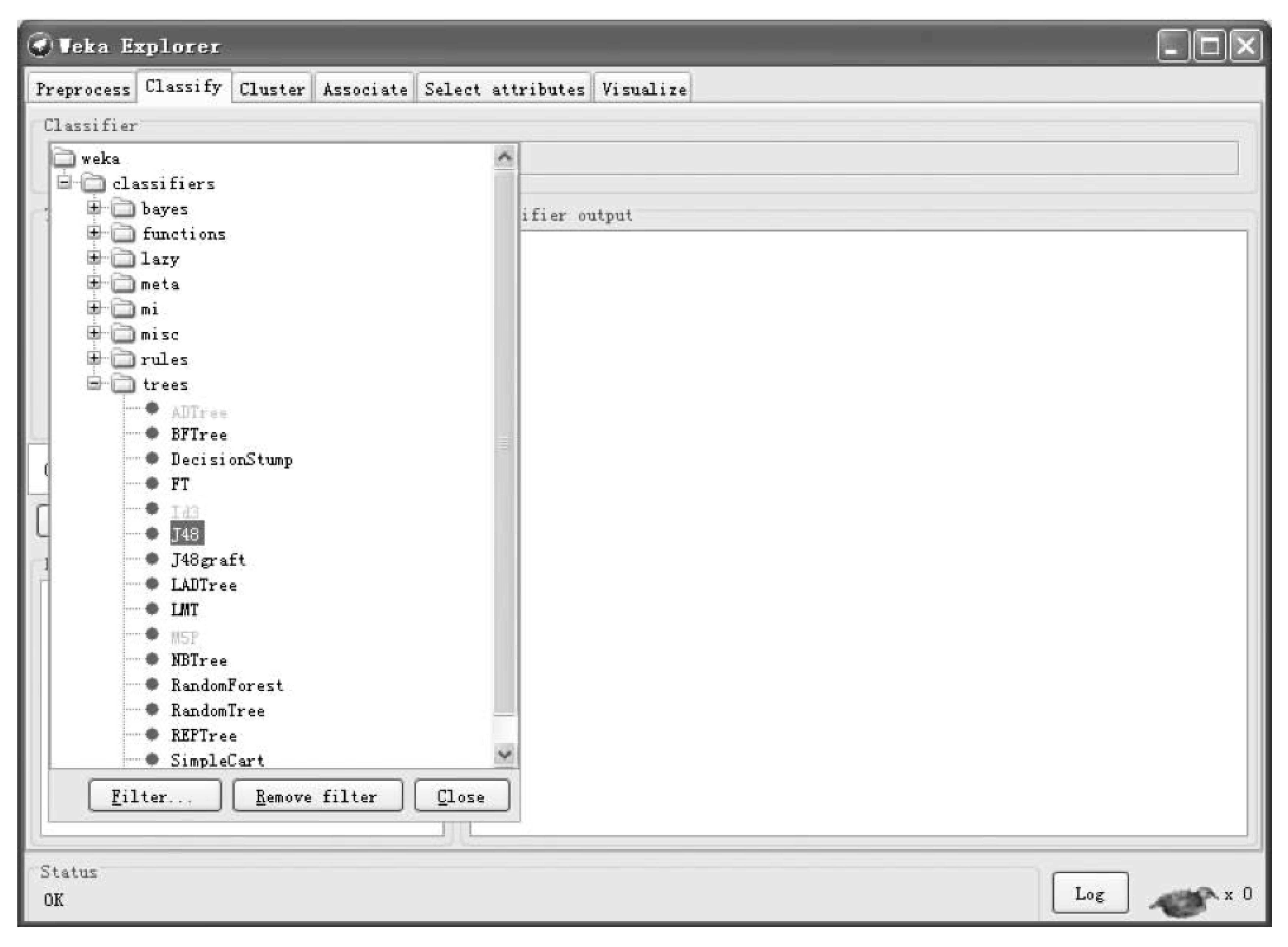


图 10-7　选择分类方法

(12) 选择 Test options 选项中的 Supplied test set 单选按钮，单击 Set 按钮，如图 10-8 所示。

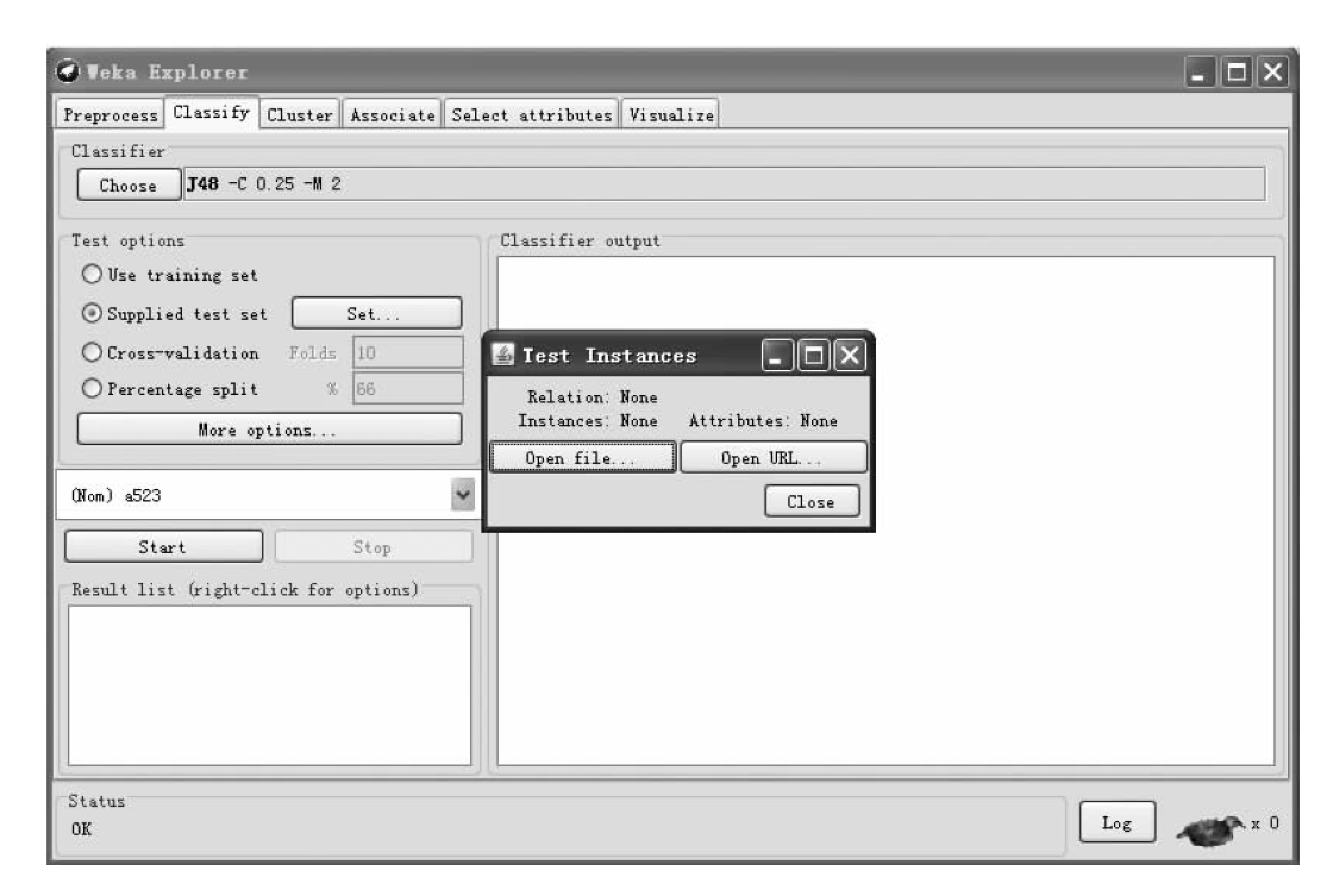


图 10-8 设置测试选项

(13) 选择"文本特征向量矩阵 30.csv"文件，并单击"打开"按钮，如图 10-9 所示。



图 10-9 打开测试数据文件

(14) 单击 Start 按钮，Weka 软件显示运行结果，如图 10-10 所示。从结果可以看出，30 篇文本中 28 篇文本得到了正确的分类结果，正确率为 93.33%。

(15) 右击 Result list 中刚才出现的那一项，在弹出的菜单中选择 Visualize tree 项，如图 10-11 所示。

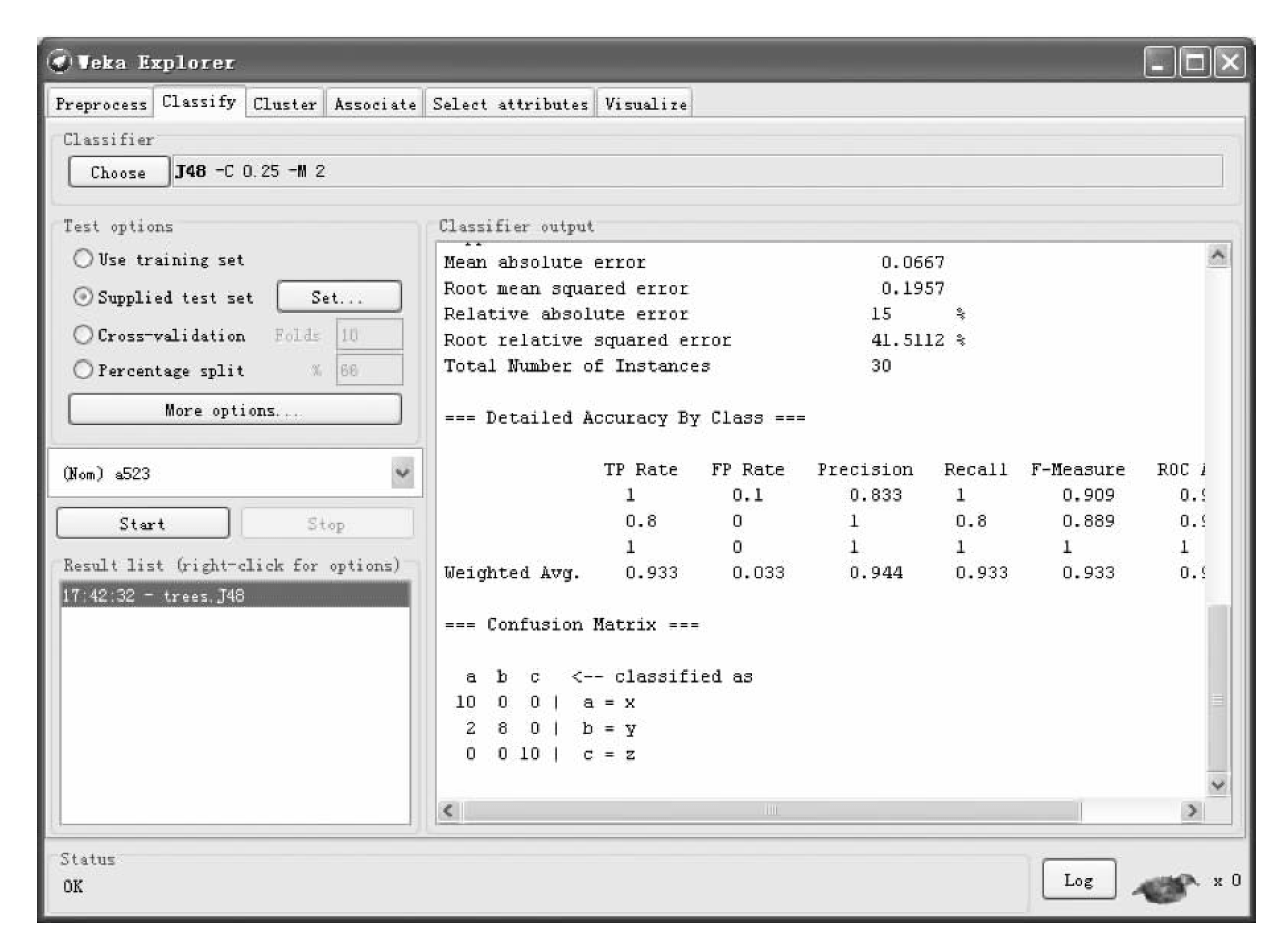


图 10-10　运行分类方法

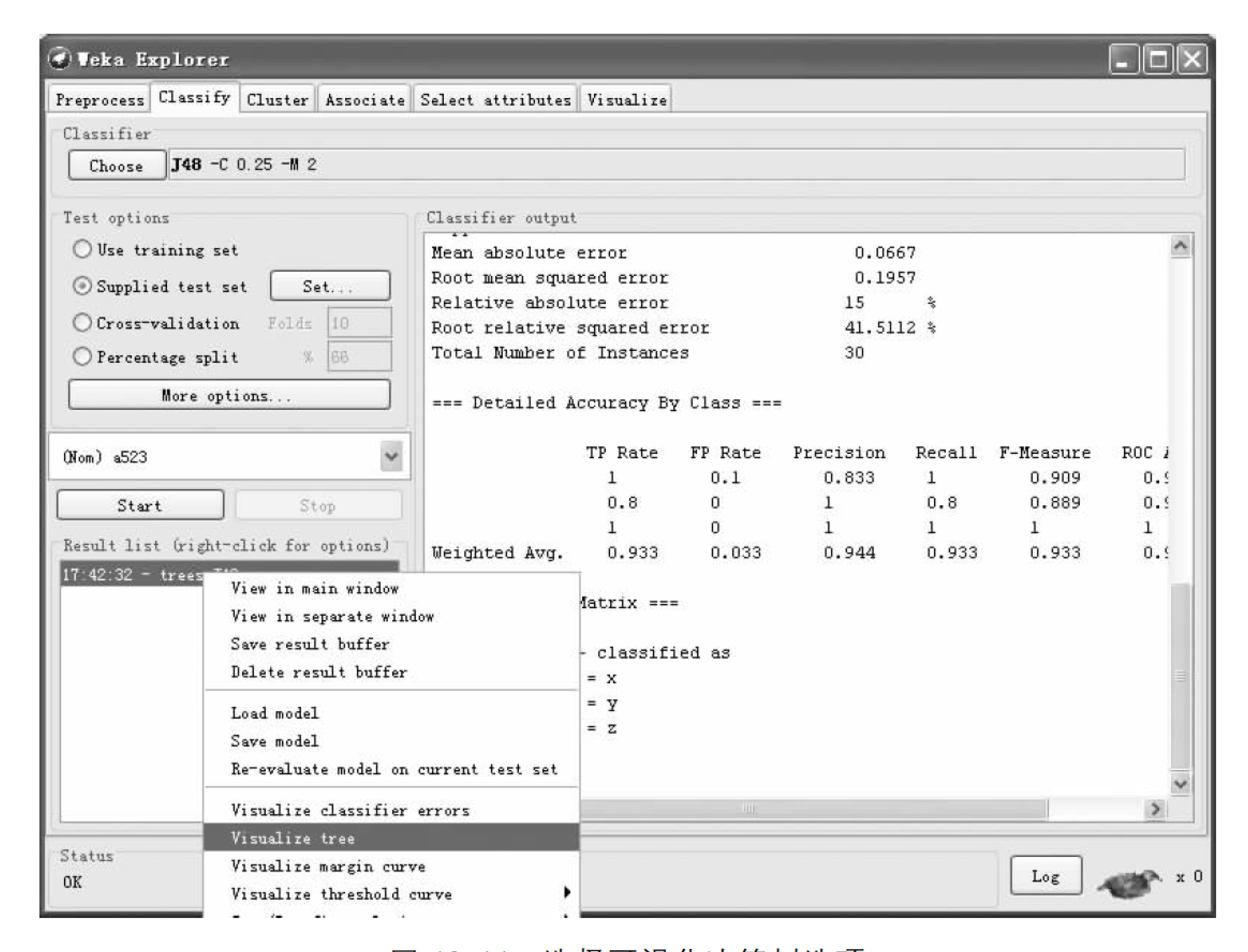


图 10-11　选择可视化决策树选项

（16）新窗口中可以看到图形模式的决策树，如图 10-12 所示。

（17）右击 Result list 中刚才出现的那一项，在弹出的菜单中选择 Visualize classifier error 项，如图 10-13 所示。

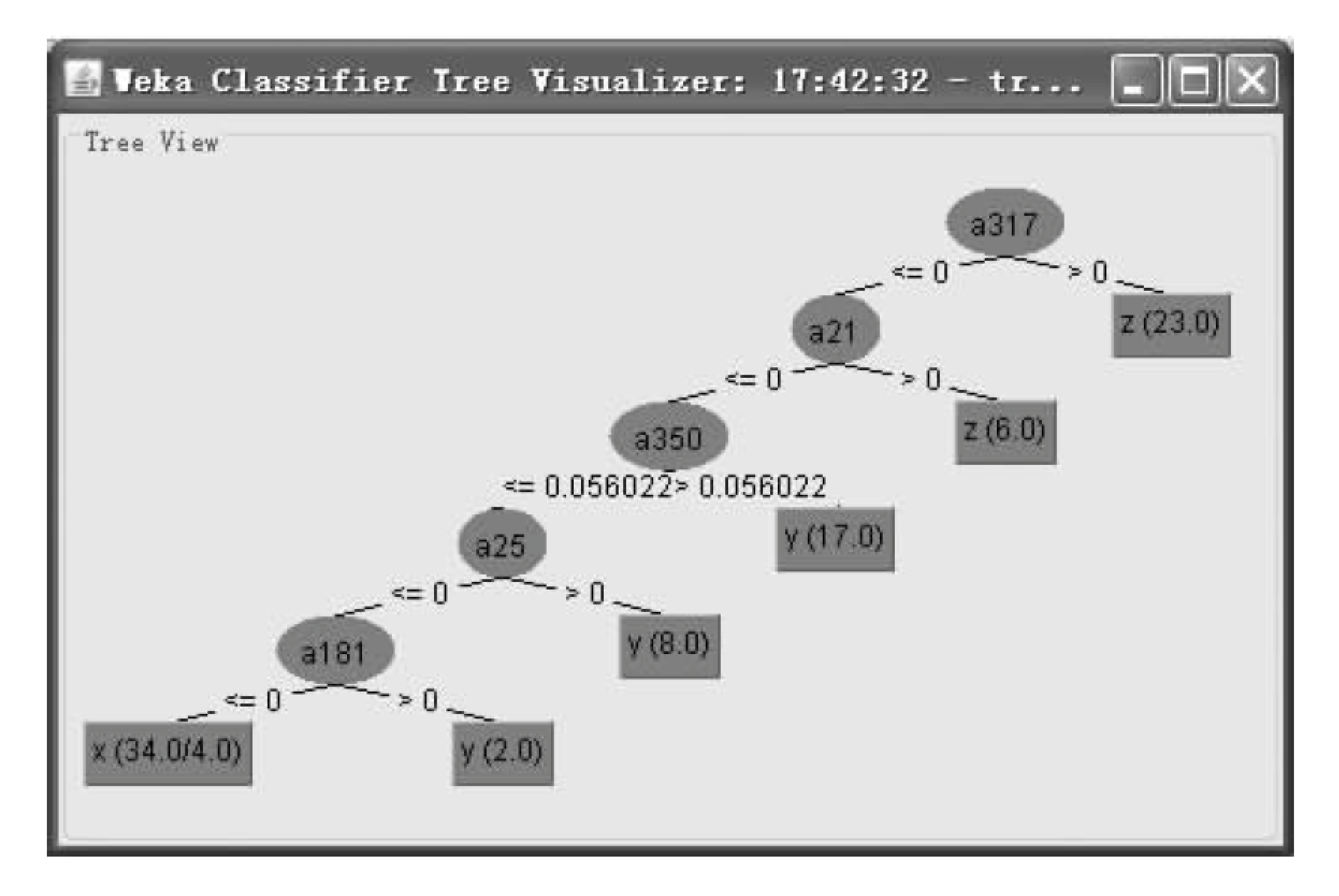


图 10-12　查看决策树

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose　J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set　Set...

Cross-validation　Folds 10

Percentage split　% 66

More options...

(Nom) a523

Start　Stop

Result list (right-click for options)

15:14:13 - rules.ZeroR

15:14:46 - trees.J48

View in main window
View in separate window
Save result buffer
Delete result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifier errors
Visualize tree
Visualize margin curve

Classifier output

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          28               93.3333 %
Incorrectly Classified Instances         2                6.6667 %
Kappa statistic                          0.9
Mean absolute error                      0.0667
Root mean squared error                  0.1957
Relative absolute error                 15      %
Root relative squared error             41.5112 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
                   1         0.1        0.833     1         0.909     0.9
                   0.8       0          1         0.8       0.889     0.9
                   1         0          1         1         1         1
                   0.933     0.033      0.944     0.933     0.933     0.9

atrix ===

 classified as
```

Status

OK

Log　x 0

图 10-13　选择 Visualize classifier errors 选项

（18）在弹出的对话框中单击 Save 按钮，并保存为文件“中文分类详细结果.arff”，如图 10-14 和图 10-15 所示。

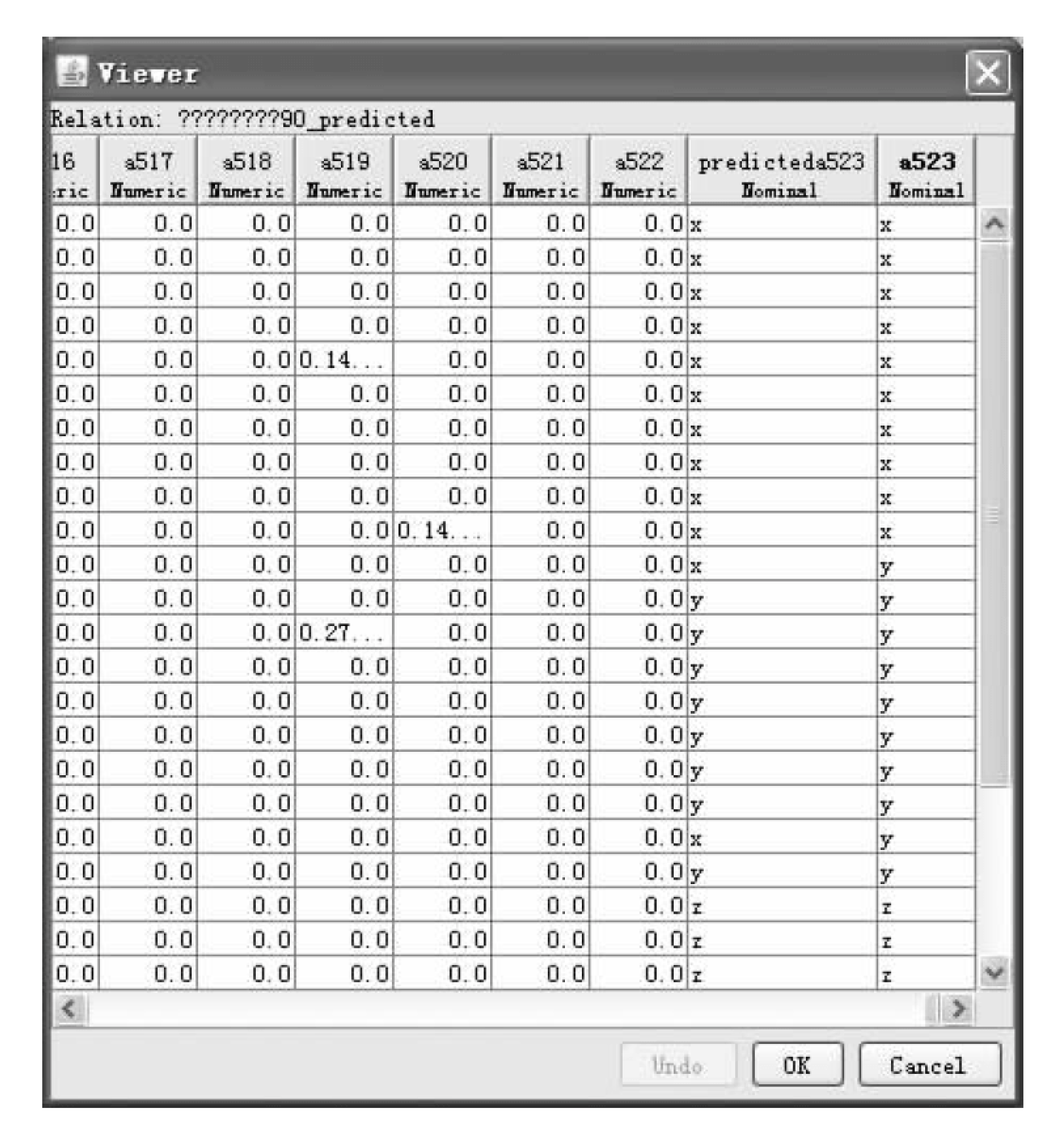
（19）打开文件“中文分类详细结果.arff”，可以查看每篇文本的实际类别和应用文本挖掘方法得到的预测类别。从结果中可以看出有两篇电影类别的文本被预测成了军事类别文本，如图 10-16 所示。



图 10-14　查看可视化分类结果



图 10-15　将分类结果保存于文件


Viewer

Relation: ?????????90_predicted

16 ric	a517 Numeric	a518 Numeric	a519 Numeric	a520 Numeric	a521 Numeric	a522 Numeric	predicteda523 Nominal	a523 Nominal
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.14...	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.14...	0.0	0.0	x	x
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.27...	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	x	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	y	y
0.0	0.0	0.0	0.0	0.0	0.0	0.0	z	z
0.0	0.0	0.0	0.0	0.0	0.0	0.0	z	z
0.0	0.0	0.0	0.0	0.0	0.0	0.0	z	z

Undo OK Cancel

图 10-16　查看预测结果

10.4　案例小结

随着信息技术的发展，文本数据的数量急剧增长，所以就有必要实现对文本数据的自动挖掘。本案例利用由 90 篇文本组成的训练集，采用决策树方法构建了分类器，最终实现了对 30 篇文本的自动分类，并取得了较高的正确率。对中文文本实现自动分类的关键是对文本进行预处理：首先，进行分词，这一步骤可以借用现成的软件进行实现；然后，在词频矩阵的基础上利用 TFIDF 公式得到文本的特征向量矩阵；最后，可以采用数据挖掘中的分类方法实现对文本的分类。在案例的实现过程中，文本特征向量矩阵是一个高维矩阵，通过属性选取方法可以使得实现过程进一步得到优化，也可以试着采用数据挖掘的其他分类方法看一下分类结果的正确率。

附录 A　SQL Server 2005 的安装

A1　任务描述

凭借全面的功能和高度的集成性，以及对日常任务的自动化管理能力，SQL Server 2005 为不同需求的用户提供了一个可靠、安全和高效的平台，用于数据管理、数据挖掘和商务智能。安装 SQL Server 2005 基本上与其他 Windows 产品类似，与 SQL Server 以前版本的区别在于要有.NET 框架的支持。

请操作实现 SQL Server 2005 软件在 Windows XP 系统下的安装过程。

A2　具体实现

（1）双击 SQL Server 2005 安装软件包中的可执行程序 setup.exe，如图 A-1 所示。



图 A-1　选择可执行程序

（2）在弹出的“Microsoft SQL Server 2005 安装程序”对话框中，选中“我接受许可条款和条件”复选框，单击“下一步”按钮，如图 A-2 所示。

（3）在弹出的“安装必备组件”页面中，单击“安装”按钮，如图 A-3 所示。

（4）安装成功后，单击“下一步”按钮，在弹出的“欢迎使用 Microsoft SQL Server 安装向导”页面中，单击“下一步”按钮，如图 A-4 所示。

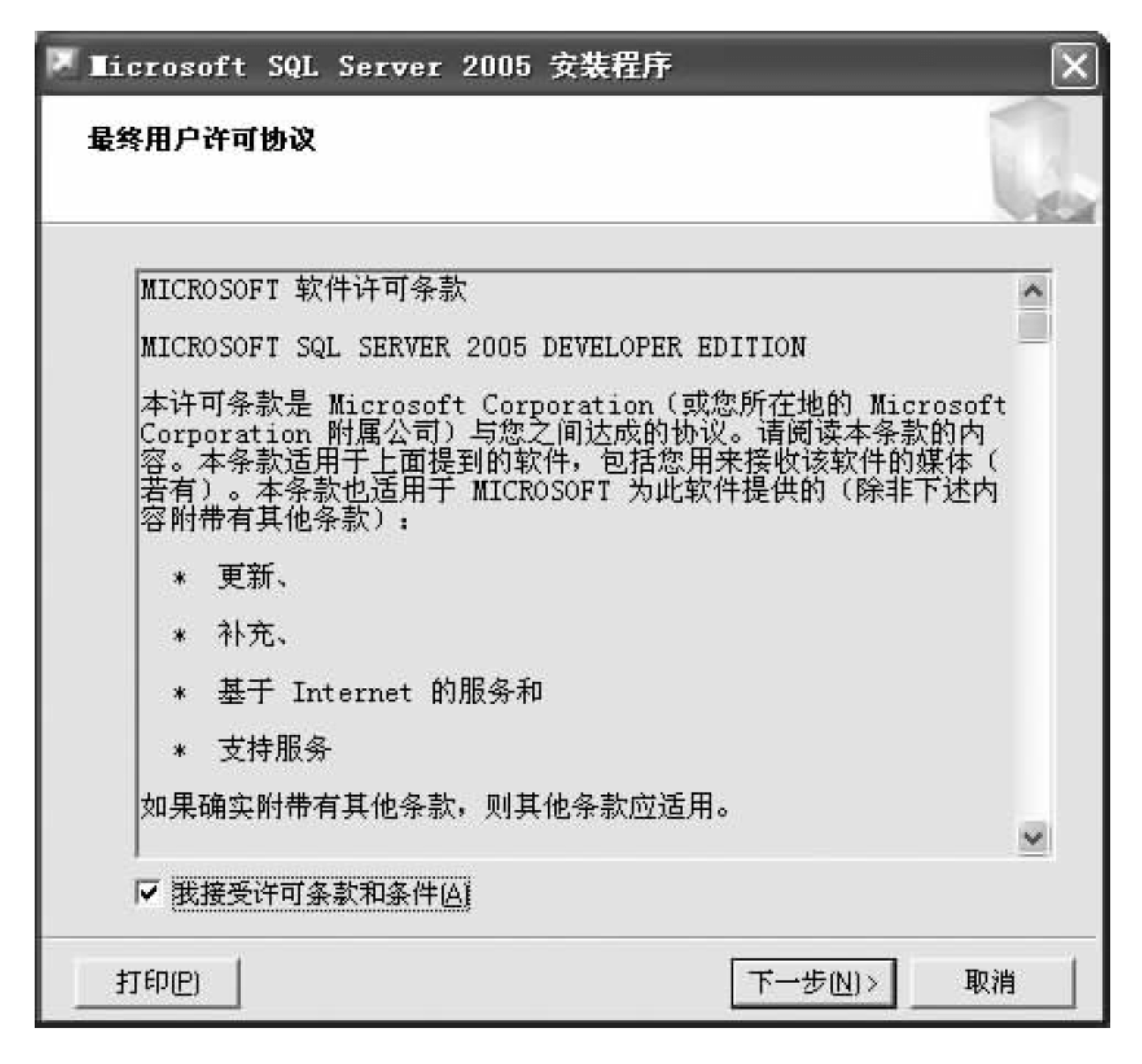


图 A-2　接受许可条款

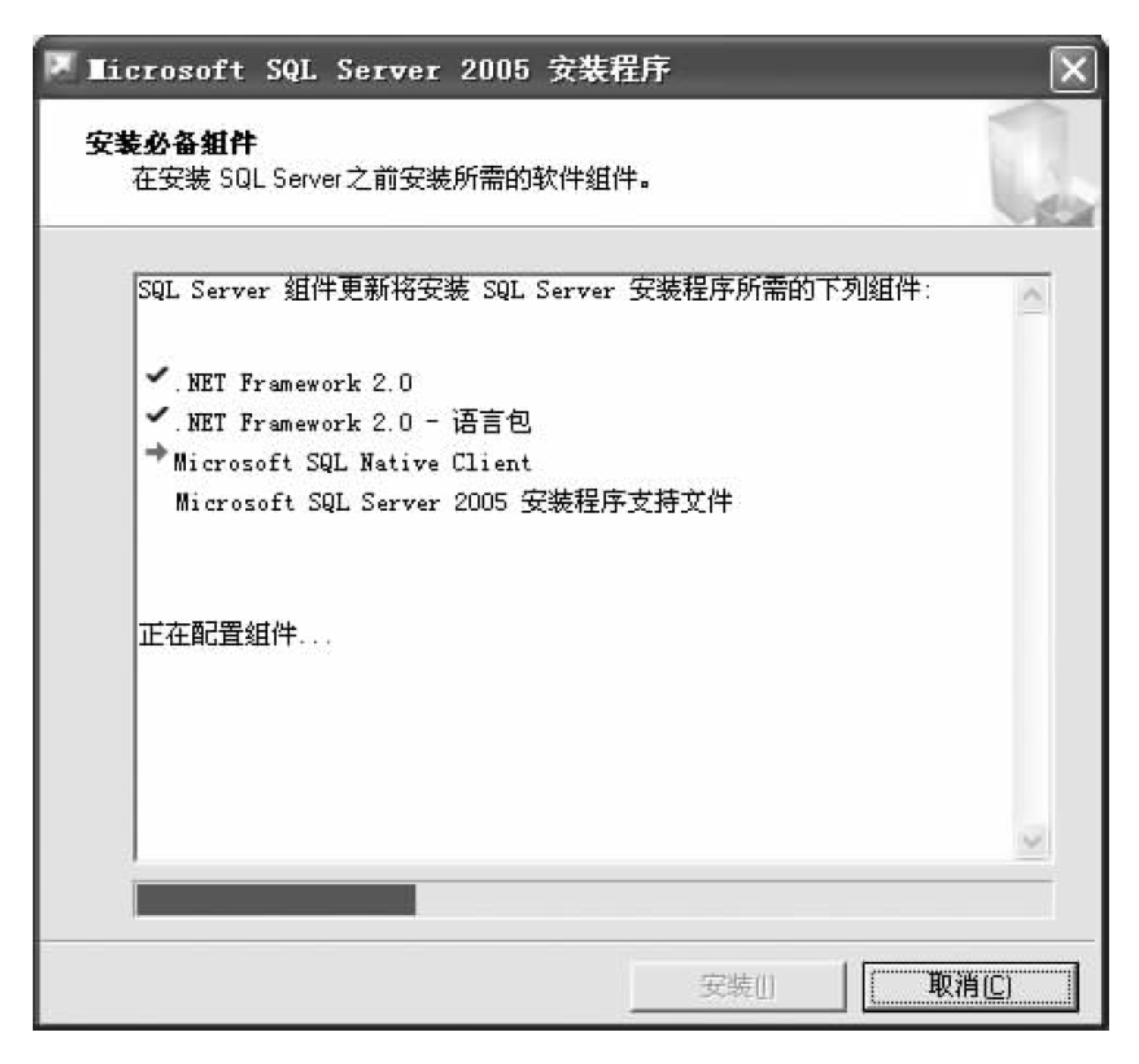


图 A-3　安装必备组件



图 A-4　使用安装向导

(5) SQL Server 将自动检查系统中可能存在的潜在问题，最常见的问题是，系统提示“IIS 未安装或未启用”，如果安装了 IIS 服务请启动 IIS 服务，如果没有安装请安装 IIS 服务。系统配置成功后，单击“下一步”按钮，如图 A-5 所示。



图 A-5　系统配置检查

(6) 注册信息填写姓名和 25 个字符的产品密钥，单击“下一步”按钮，如图 A-6 所示。

(7) 选择需要安装的组件，单击“下一步”按钮，如图 A-7 所示。

① SQL Server Database Services：SQL Server 默认选中的基础服务。

② Analysis Services：在 SQL Server 2005 中，分析服务(Analysis Services)提供了一

Microsoft SQL Server 2005 (64 位)安装程序

注册信息

下列信息将对您安装的系统进行个性化设置。

在继续操作之前，必须填写"姓名"字段。"公司"字段是可选的。

姓名(A):

[redacted]

公司(O):

[redacted]

请输入 25 个字符的产品密钥。在 CD 内衬说明的黄色不干胶标签或 CD 封套上可找到该号。

[redacted]

帮助(H)　< 上一步(B)　下一步(N) >　取消

图 A-6　设置注册信息



图 A-7　选择安装的组件

个统一和集成的商业数据视图，可被用做所有传统报表、QLAP 分析和数据挖掘的基础（在这里选中安装）。

③ Reporting Services：SQL Server 2005 Reporting Services 是一个基于服务器的企业级报表环境，可借助 web services 进行管理。通过把报表作为更进一步的商业智能的数据源来分发，复杂的分析可被更多的用户所用(在这里选中安装)。

④ Notification Services：通知服务，具有现实的商业价值，它们吸引客户，让雇主更高效，决策更灵敏。

⑤ Integration Services：SQL Server 2005 带来了一个全新的企业级数据整合平台。此平台具有出色的整合能力，使得组织机构能更加容易地管理来自于不同的关系型和非关系型数据源的数据。

(8) 选择命名实例，实例名填写 DATAMINING，单击"下一步"按钮，如图 A-8 所示。

① 默认实例：用计算机在网络上的名字来命名实例。如果应用程序在请求连接 SQL

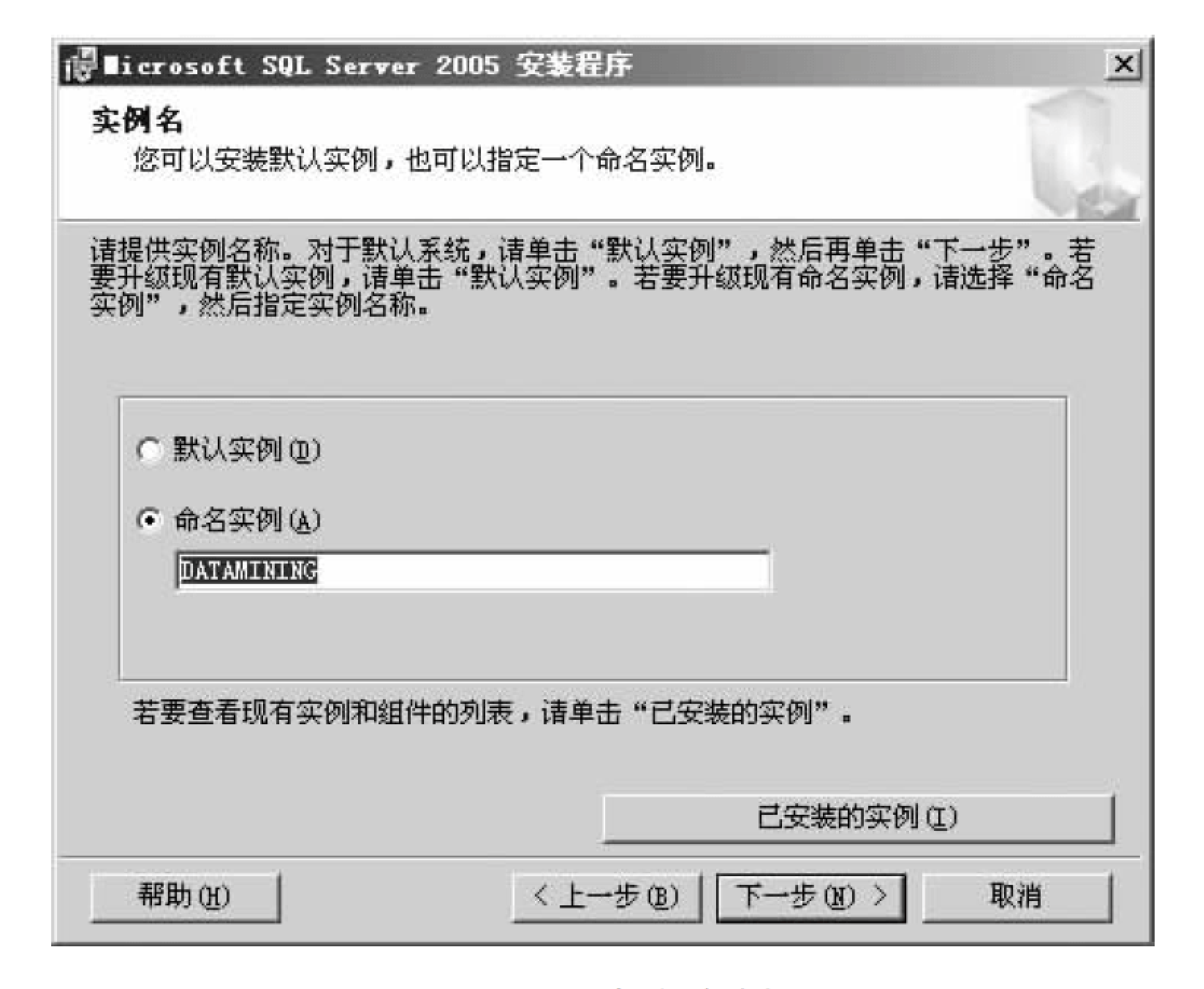


图 A-8　命名实例

Server 时只指定了计算机名，则 SQL Server 客户端组件将尝试连接这台计算机上的数据库引擎默认实例。这保留了与现有 SQL Server 应用程序的兼容性。一台计算机上只能有一个默认实例，而默认实例可以是 SQL Server 的任何版本。

② 命名实例：通过使用计算机在网络上的名字加上实例名字来进行标识的实例。就是在有了默认实例后，命名其他的实例，就需要再次安装命名实例，一台计算机可以同时拥有多个命名实例。

(9) 在服务账户中选择“使用内置系统账户”中的“本地系统”项，在“安装结束时启动服务”选项组中选中 SQL Server、Reporting Services 和 Analysis Services 复选框，单击“下一步”按钮，如图 A-9 所示。

图 A-9　设置服务账户

（10）在身份验证模式中选择“Windows 身份验证模式”单选按钮，单击“下一步”按钮，如图 A-10 所示。



图 A-10 设置身份验证模式

（11）依次单击“下一步”按钮，如图 A-11 和图 A-12 所示。



图 A-11 进行排序规划设置

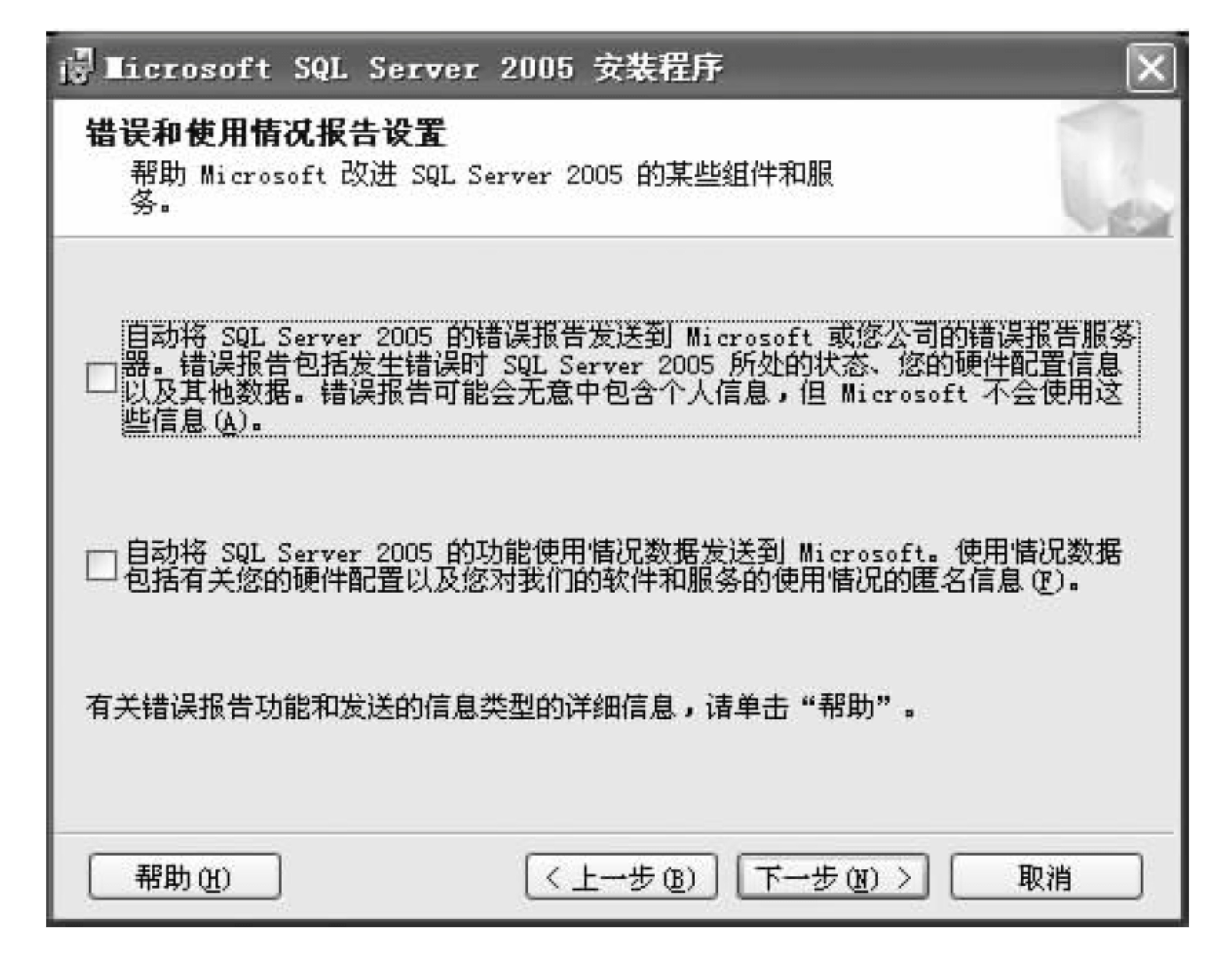


图 A-12 进行错误和使用报告设置

(12) 单击“安装”按钮，如图 A-13 所示。



图 A-13 进行安装

(13) 当安装组件完毕时如图 A-14 所示，单击“下一步”按钮，然后单击“完成”按钮，如图 A-15 所示，到此为止 SQL Server 2005 服务器程序安装完成。



图 A-14　安装进度显示

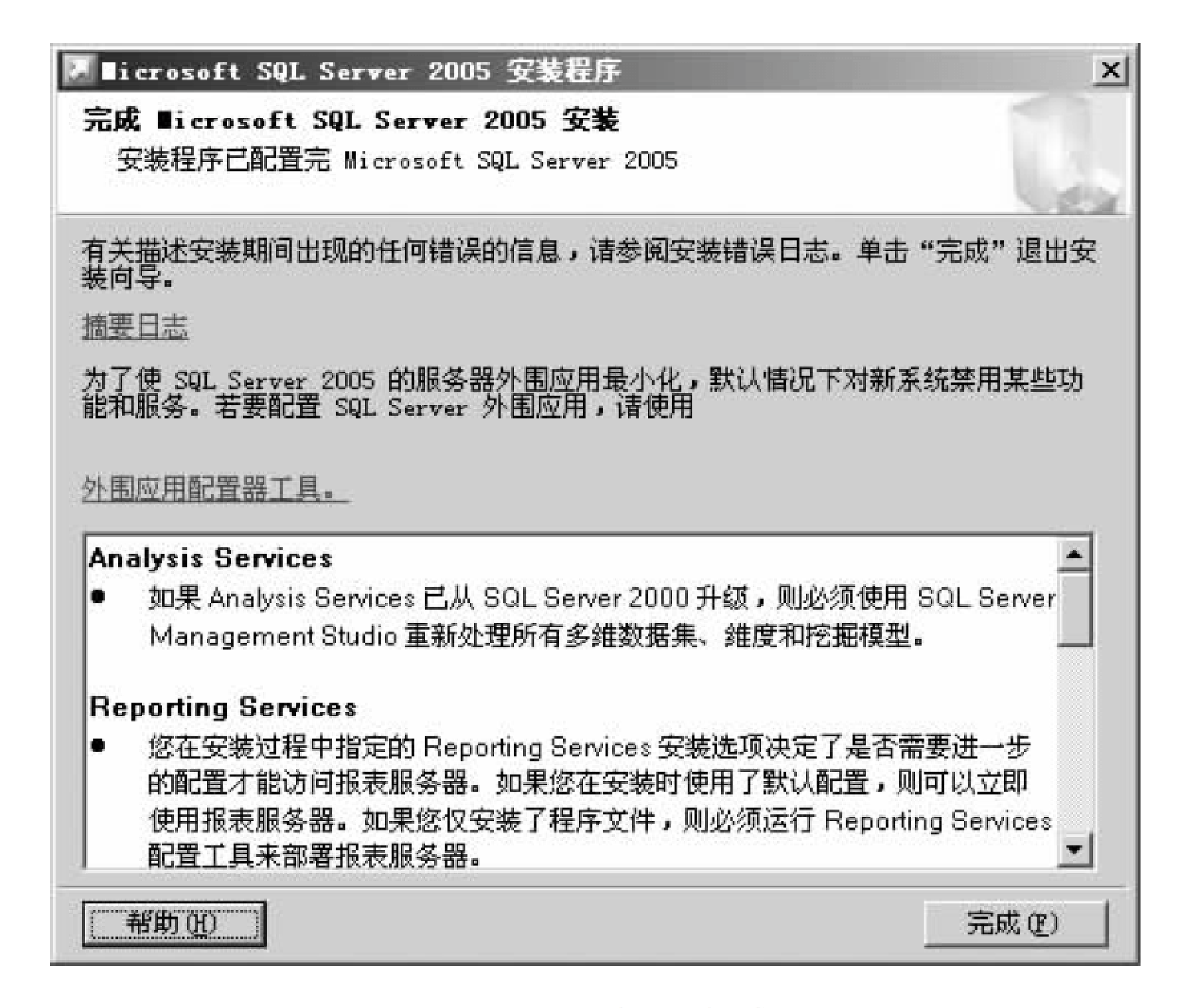


图 A-15　安装完成

附录 B　Weka 软件的安装和数据转换

B1　任 务 描 述

Weka 的全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis),是由新西兰怀卡托大学开发的专业数据挖掘系统,为数据挖掘应用提供了一个统一的界面,可以用许多不同的学习算法处理给定的任何数据集,并能够评估不同的学习算法所得出的结果。2005 年 8 月,在第 11 届 ACM SIGKDD 国际会议上,怀卡托大学的 Weka 小组荣获了数据挖掘和知识探索领域的最高服务奖,Weka 系统得到了广泛的认可,被誉为数据挖掘和机器学习历史上的里程碑,是现今最完备的数据挖掘工具之一。在本案例中,试完成以下任务:

(1) 下载并安装 Weka 软件。

(2) 如何查看 ARFF 格式的文件。

(3) 如何将数据集转换成 ARFF 格式。

B2　具 体 实 现

1. 下载并安装 Weka 软件

(1) 输入网址 http://www.cs.waikato.ac.nz/ml/weka,进入 Weka 软件下载页面,并点击“Download”选项,如图 B-1 所示。



图 B-1　进入 Weka 软件下载页面

(2) 选中要下载的软件版本,进行下载,并保存,如图 B-2 和图 B-3 所示。

(3) 单击下载的 Weka 软件,开始进行软件的安装,如图 B-4 所示。

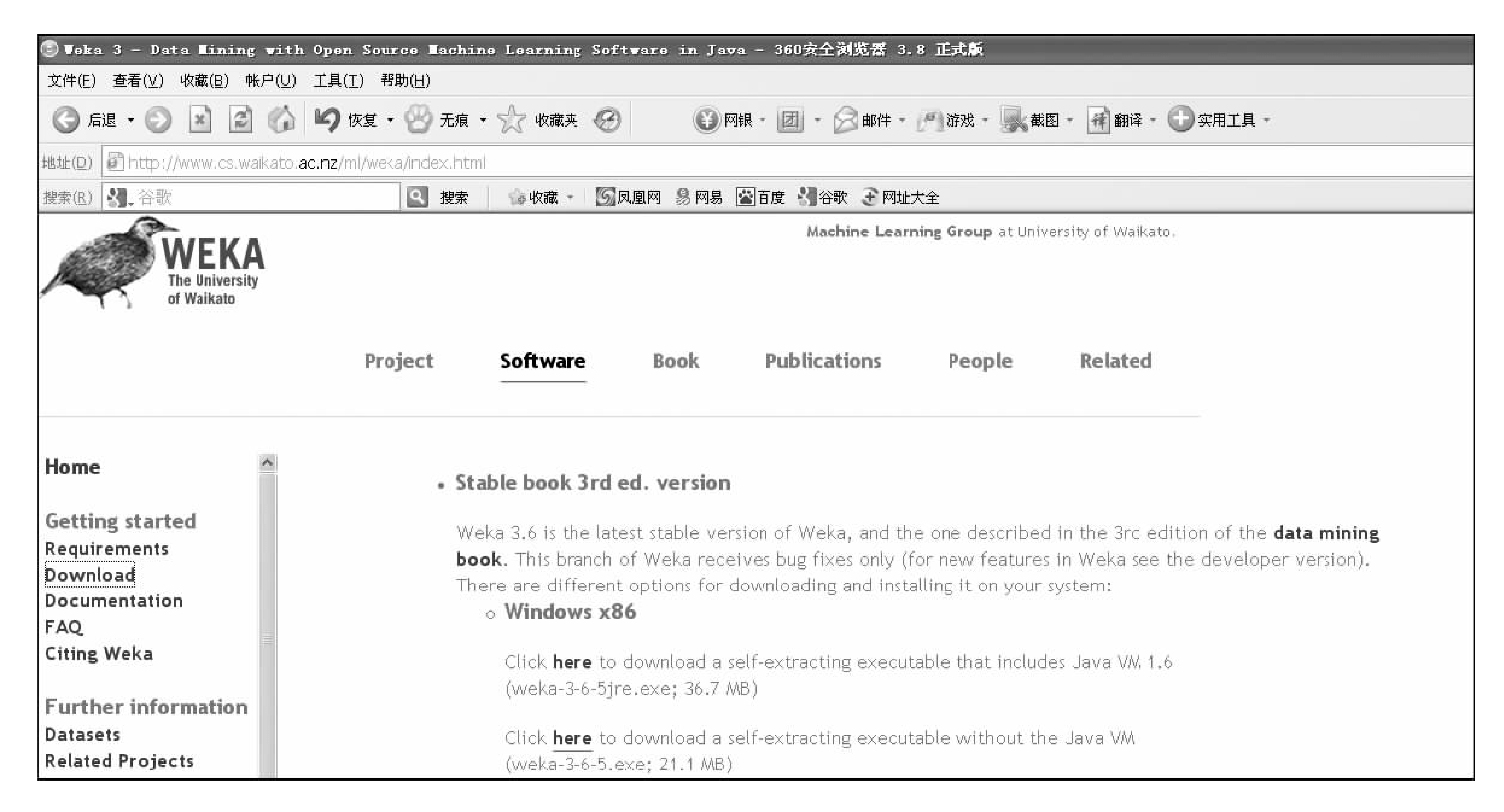


图 B-2　选择软件版本

图 B-3　进行下载



图 B-4　开始安装软件

（4）进入安装引导界面，单击 Next 按钮，如图 B-5 所示。

（5）进入协议界面，单击 I Agree 按钮，如图 B-6 所示。



图 B-5　安装引导界面

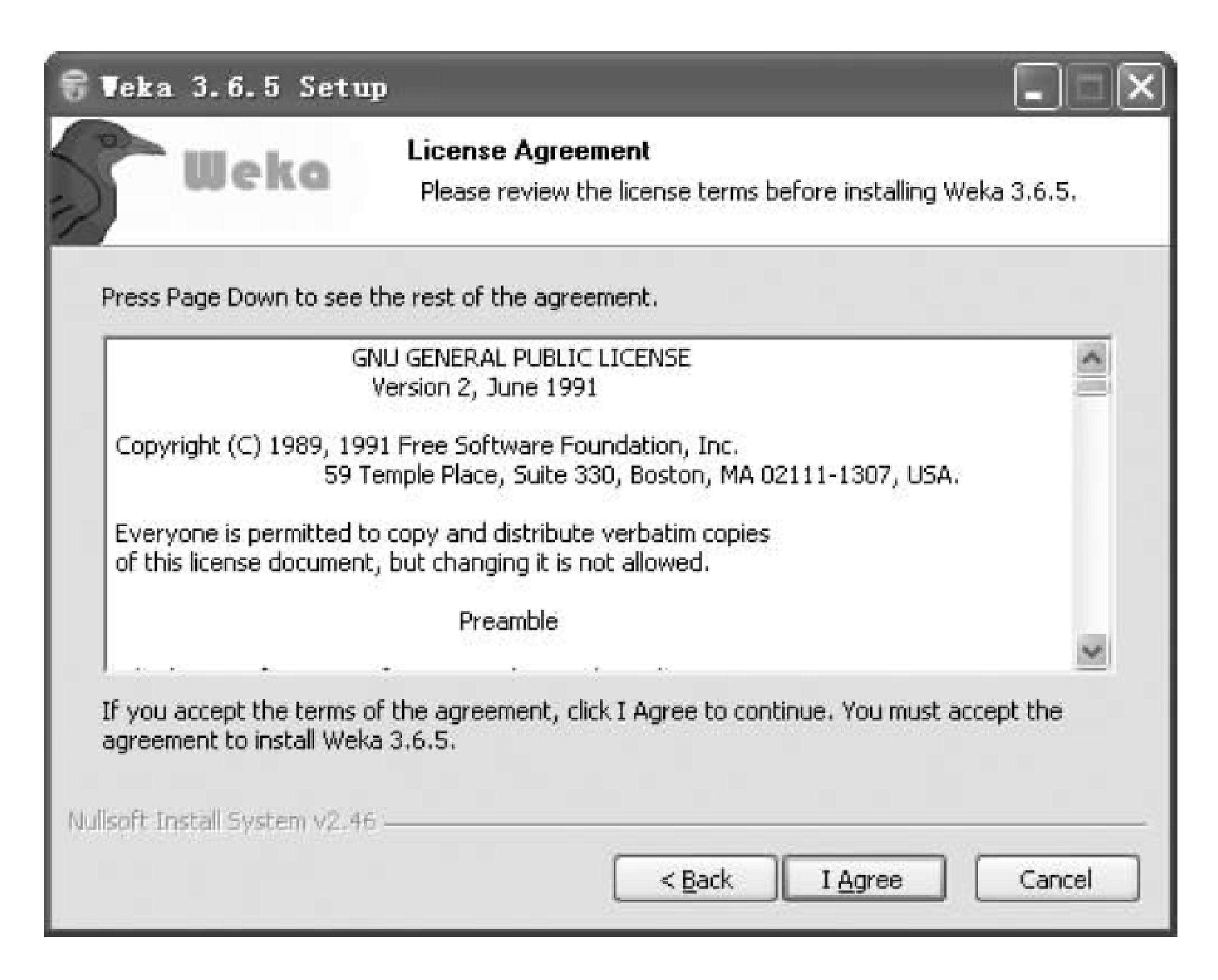


图 B-6　协议界面

（6）进入选装组件界面，选择全部安装，单击 Next 按钮，如图 B-7 所示。

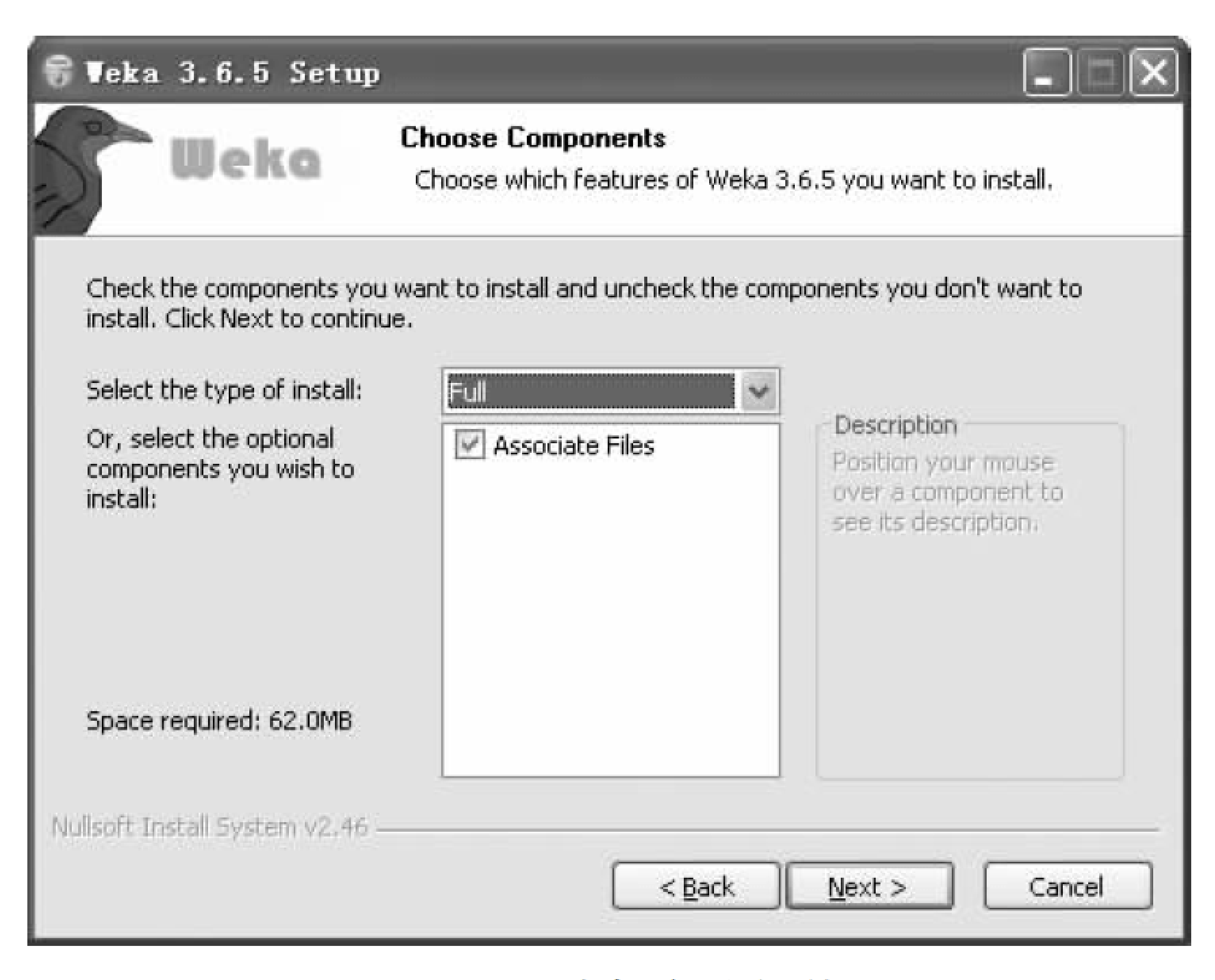


图 B-7　选择安装组件

(7) 连续单击 Next 按钮,完成 Weka 软件的安装,如图 B-8～图 B-10 所示。

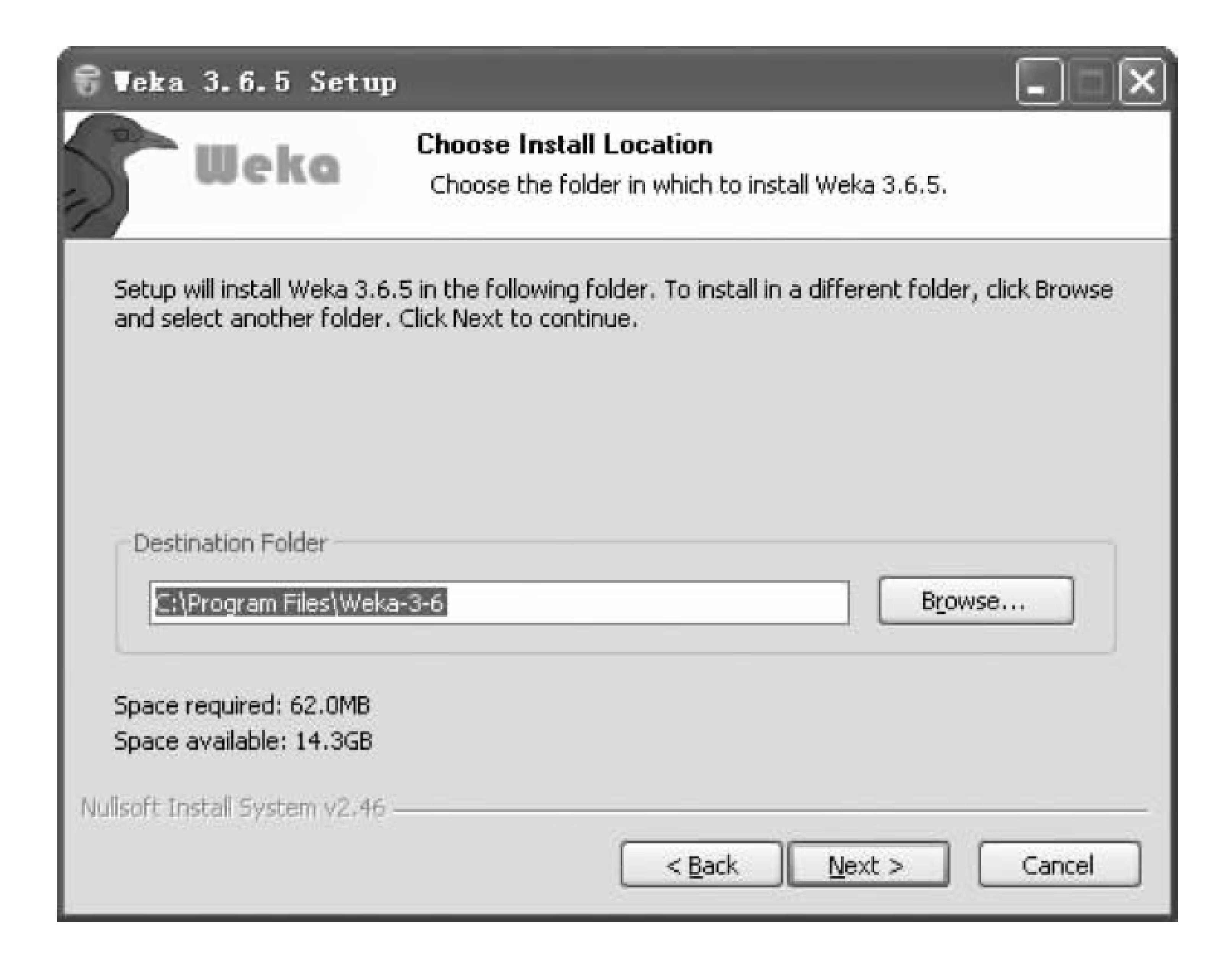


图 B-8　选择目标地址

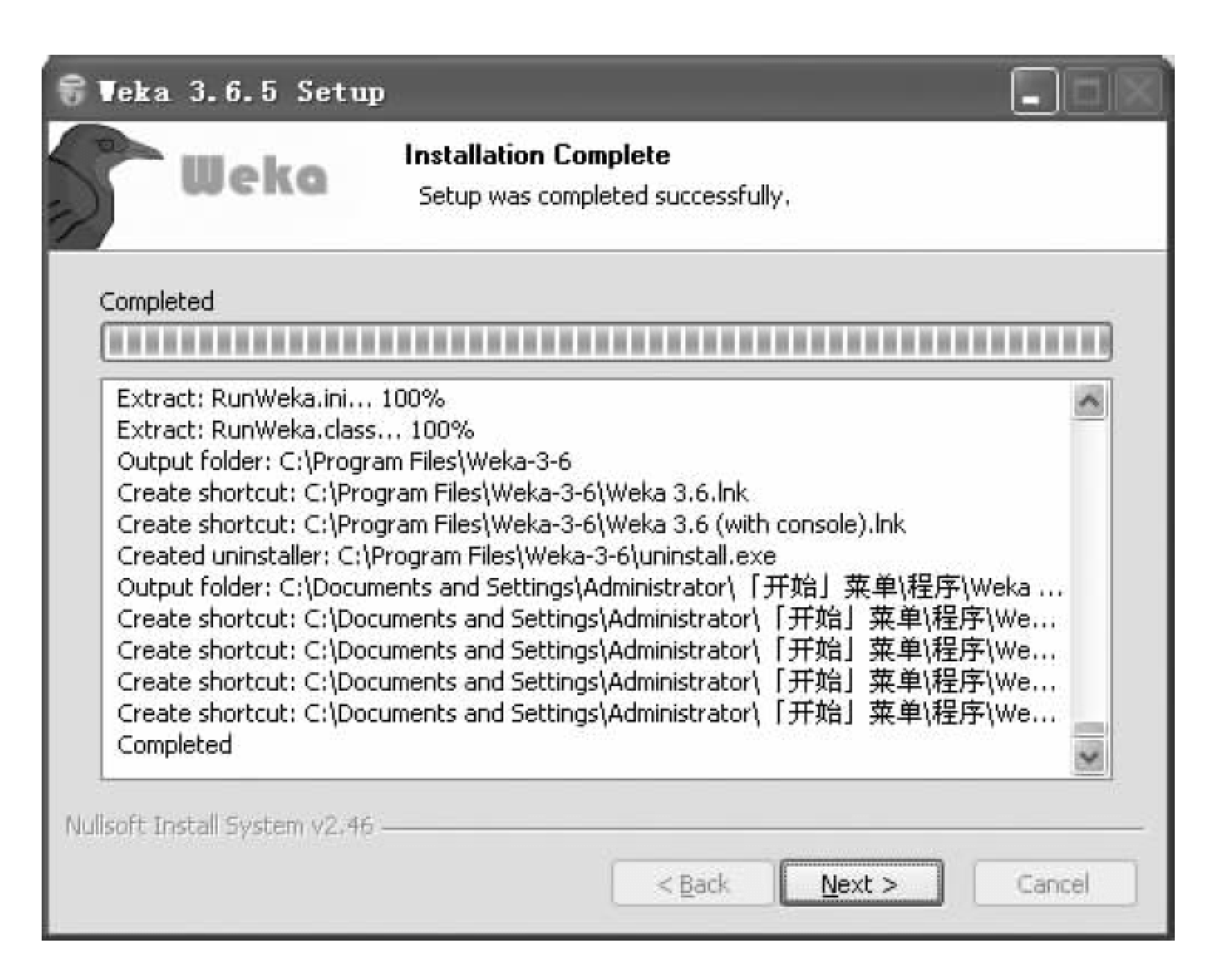


图 B-9　执行安装



图 B-10　安装完毕

2. 如何查看 ARFF 文件

Weka 对输入的数据格式有自己特殊的要求，即必须符合 ARFF（Attribute-Relation File Format）格式。ARFF 格式是 Weka 定义的一种特殊文件格式，是一种 ASCII 文本文件。Weka 自带的所有数据集都是以这种格式组织的。安装 Weka 之后，可以再安装目录中找到这些示例数据集。

（1）进入 Weka 软件的安装目录，打开 data 文件夹，如图 B-11 所示。

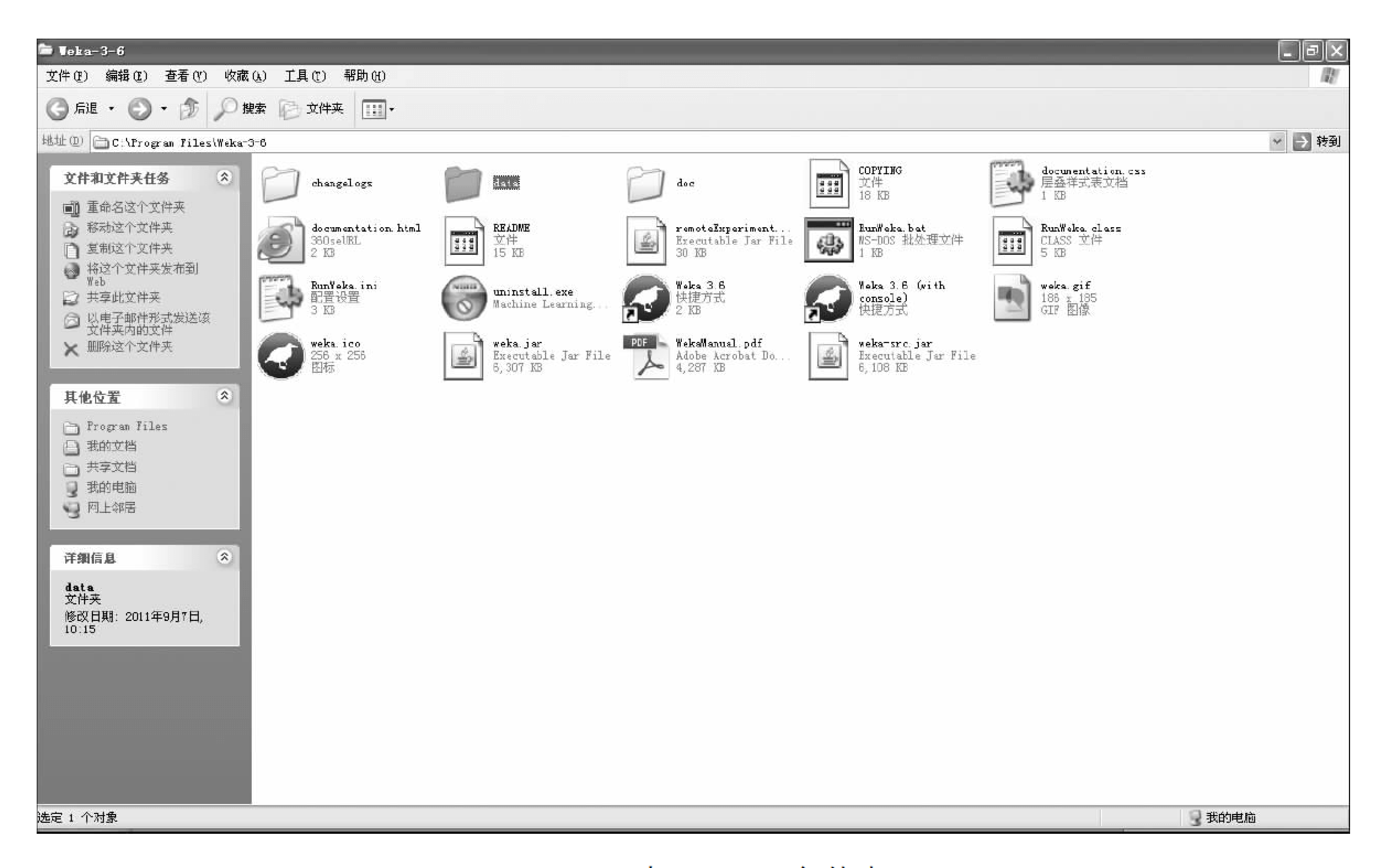


图 B-11　打开 data 文件夹

（2）选择 weather.arff 文件，在右键菜单中依次选择“打开方式”→Microsoft Office Word 命令，如图 B-12 所示。

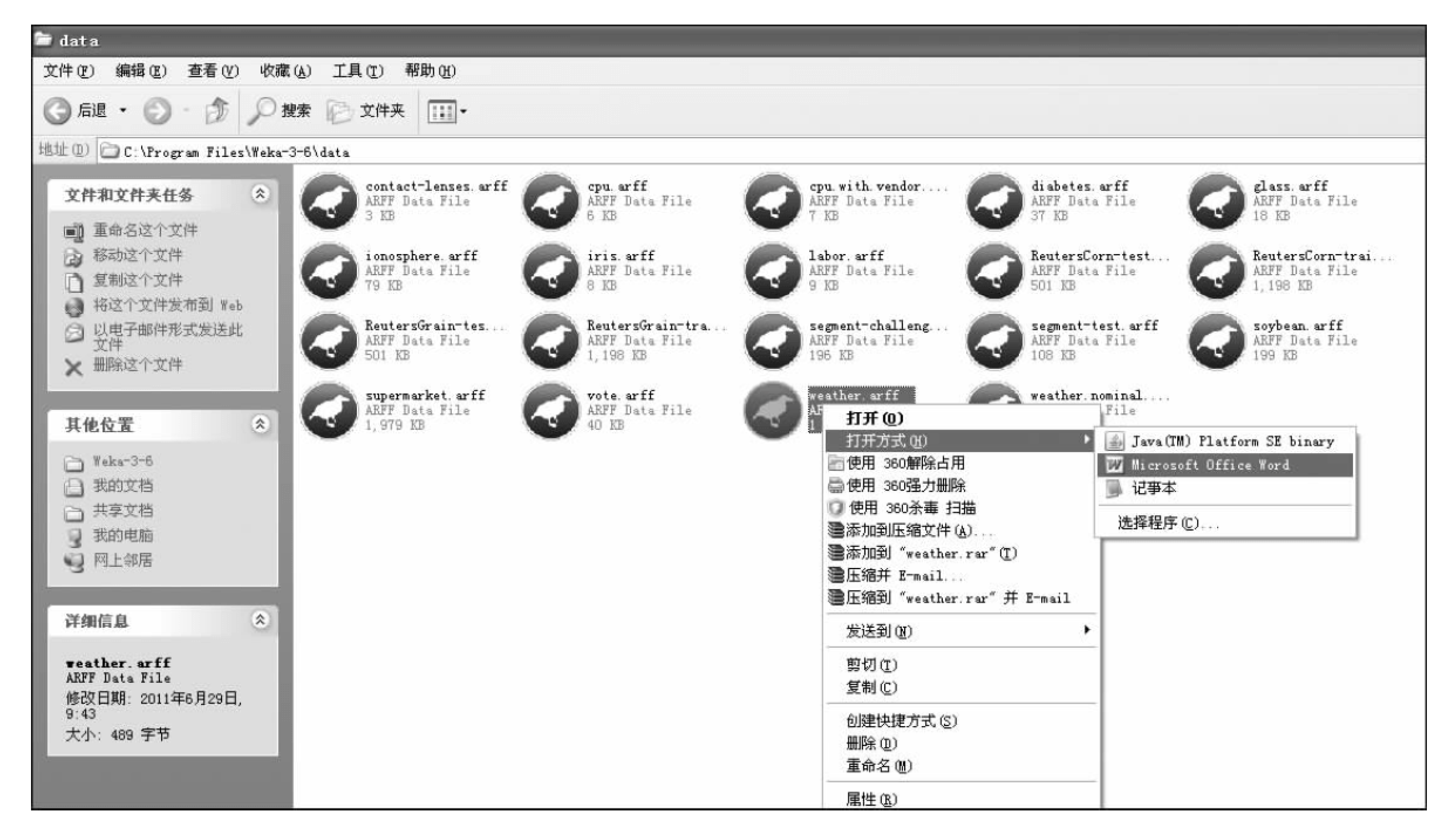


图 B-12　选择打开方式

(3) 打开文件后，文件如图 B-13 所示。跟很多数据分析软件一样，Weka 所处理的数据集是一个二维的表格。表格里的一个横行称作一个实例(Instance)，相当于统计学中的一个样本，或数据库中的一条记录。竖行称作一个属性(Attribute)，相当于统计学中的一个变量，或数据库中的一个字段。weather 文件中有 5 个属性，14 个实例。

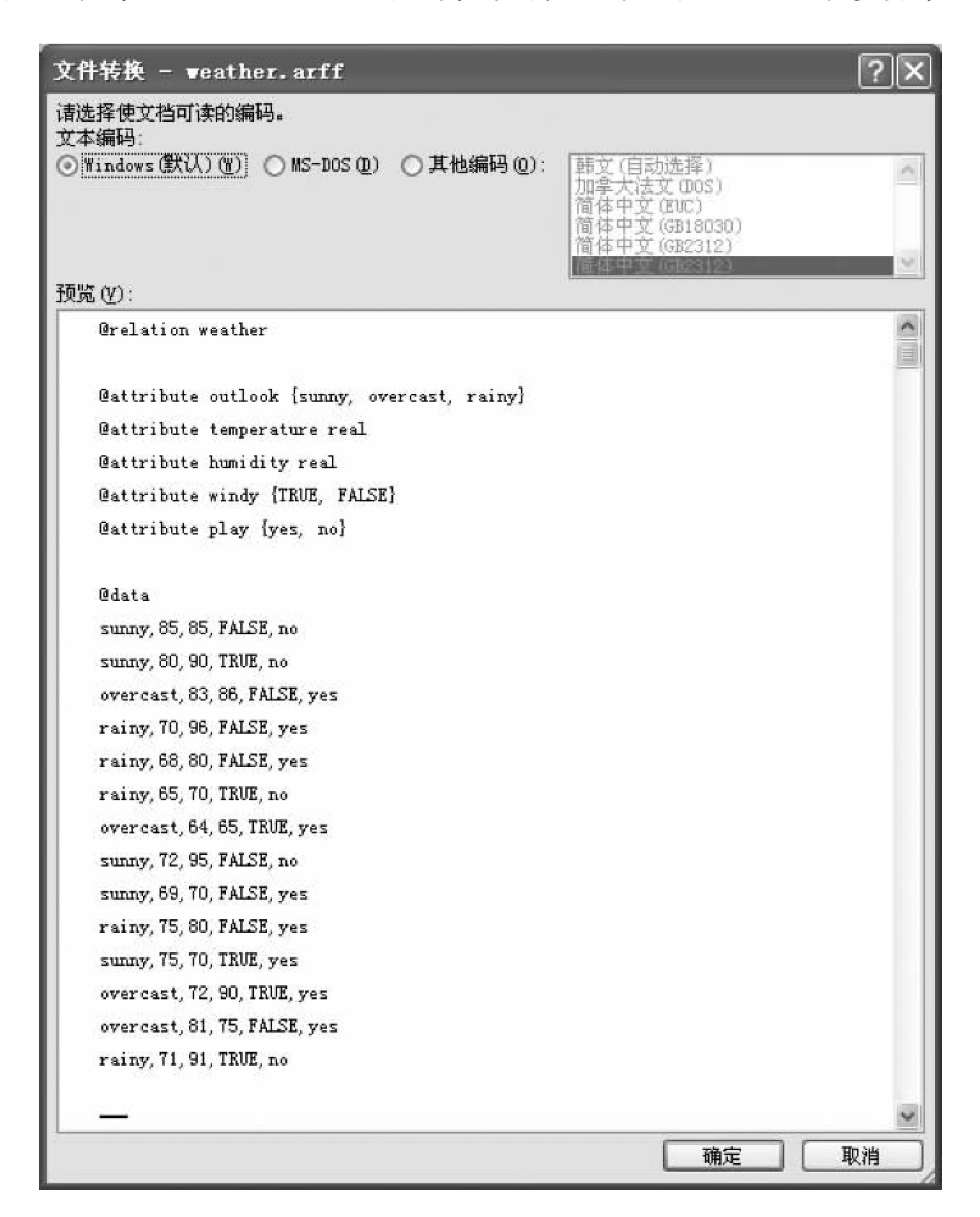


图 B-13　文件格式示例

文件中的空行将被忽略。以“%”开始的行是注释，Weka 也将忽略这些行。如果看到的 ARFF 文件多了或少了一些以“%”开始的行，这些对文件是没有影响的。

除注释之外，整个 ARFF 文件可以分为两个部分。第一部分给出了头信息(Head Information)包括了对关系的声明和对属性的声明。第二部分给出了数据信息(Data Information)，即数据集中给出的数据。从@data 标记开始，后面的就是数据信息了。

关系声明：关系名称在 ARFF 文件的第一个有效行来定义，格式为@relation<relation-name>，<relation-name>是一个字符串。

属性声明：属性声明用一列以@attribute 开头的语句表示。数据集中的每一个属性都有它对应的@attribute 语句，来定义它的属性名称和数据类型。这些声明语句的顺序很重要。首先它表明了该项属性在数据部分的位置。例如，humidity 是第三个被声明的属性，这说明数据部分那些被逗号分开的列中，第三列数据 85 90 86 96…是相应的 humidity 值。其次，最后一个声明的属性被称作 class 属性，在分类或回归任务中，它是默认的目标变量。属性声明的格式为 @attribute <datatype> ，其中 attribute 是必须以字母开头的字符串。Weka 支持的<datatype>有 4 种，分别是数值型 numeric，分类型<nominal-specification>，字符串 string，日期和时间 date [<date-format>]。

3. 如何将数据集处理成 ARFF 格式

使用 Weka 进行数据挖掘，首先需要将关系数据库中的数据、电子表格中的数据等处理成为 ARFF 格式。Weka 也提供了对 CSV 文件的支持，而这种格式是被很多其他软件所支持的。

很多应用的数据是存放在数据库中的，如 SQL Server，从数据库中的数据获得 ARFF 文件格式需要经过两个步骤：第一步，将数据从数据库中导出成 CSV 文件；第二步，将 CSV 文件转换成 ARFF 文件。下面以一个简单的例子演示这个转换过程。

图 B-14 所示是 SQL Server 2005 学生数据库 student 中的一张学生成绩表 stud_score。

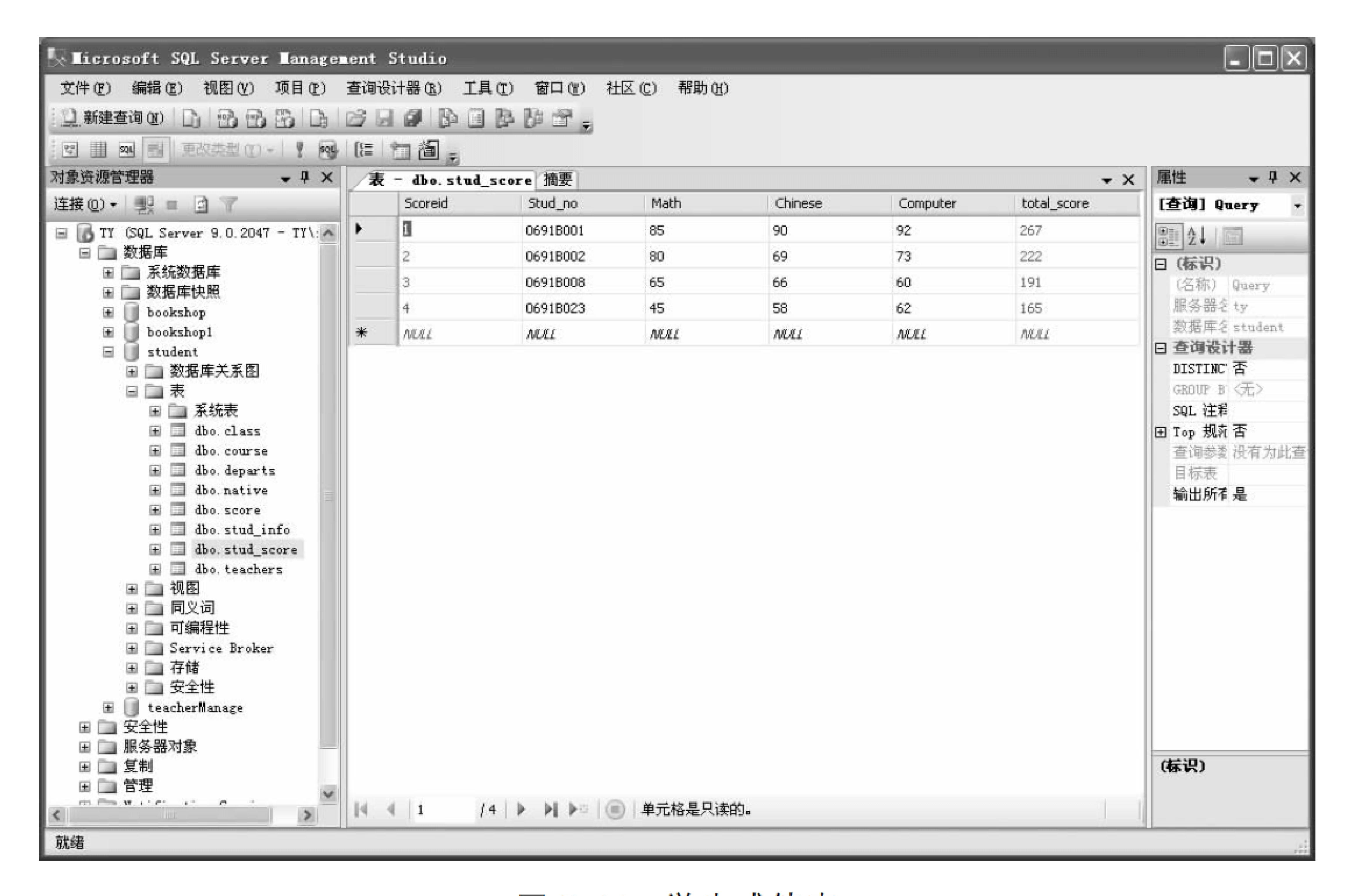


图 B-14 学生成绩表

使用 SQL Server 提供的数据导出功能，将该表导出成 CSV 文件，操作步骤如下：

(1) 点击表所属的数据库，使用右键菜单的“任务”→“导出数据”命令，如图 B-15 所示。启动数据库导入导出向导，直接单击“下一步”按钮。



图 B-15 选择导出数据命令

(2) 在图 B-16 所示的“选择数据源”页面中，选择要导出的数据库。从该对话框的“数据库”下拉列表中选择要导出的表所在的数据库。因为，在步骤(1)中已经选择了数据库，所以该对话框中默认的数据库显示为 student。可以直接单击“下一步”按钮。

图 B-16 选择数据源

(3) 从“选择目标”页面中“目标”下拉列表选择导出文件的目标格式，为了将数据导出成 CSV 格式，选择“平面文件目标”，并单击“下一步”按钮，如图 B-17 所示。



图 B-17 设置选择目标类型

(4) 单击“浏览”按钮，选择磁盘上一个已经存在 CSV 文件，如图 B-18、图 B-19 所示。如果还没有创建用来保存导出数据的 CSV 文件，则需要首先创建。或在单击“浏览”按钮打

开选择文件对话框后，在该对话框中立即创建一个 CSV 文件。



图 B-18 浏览目标页面

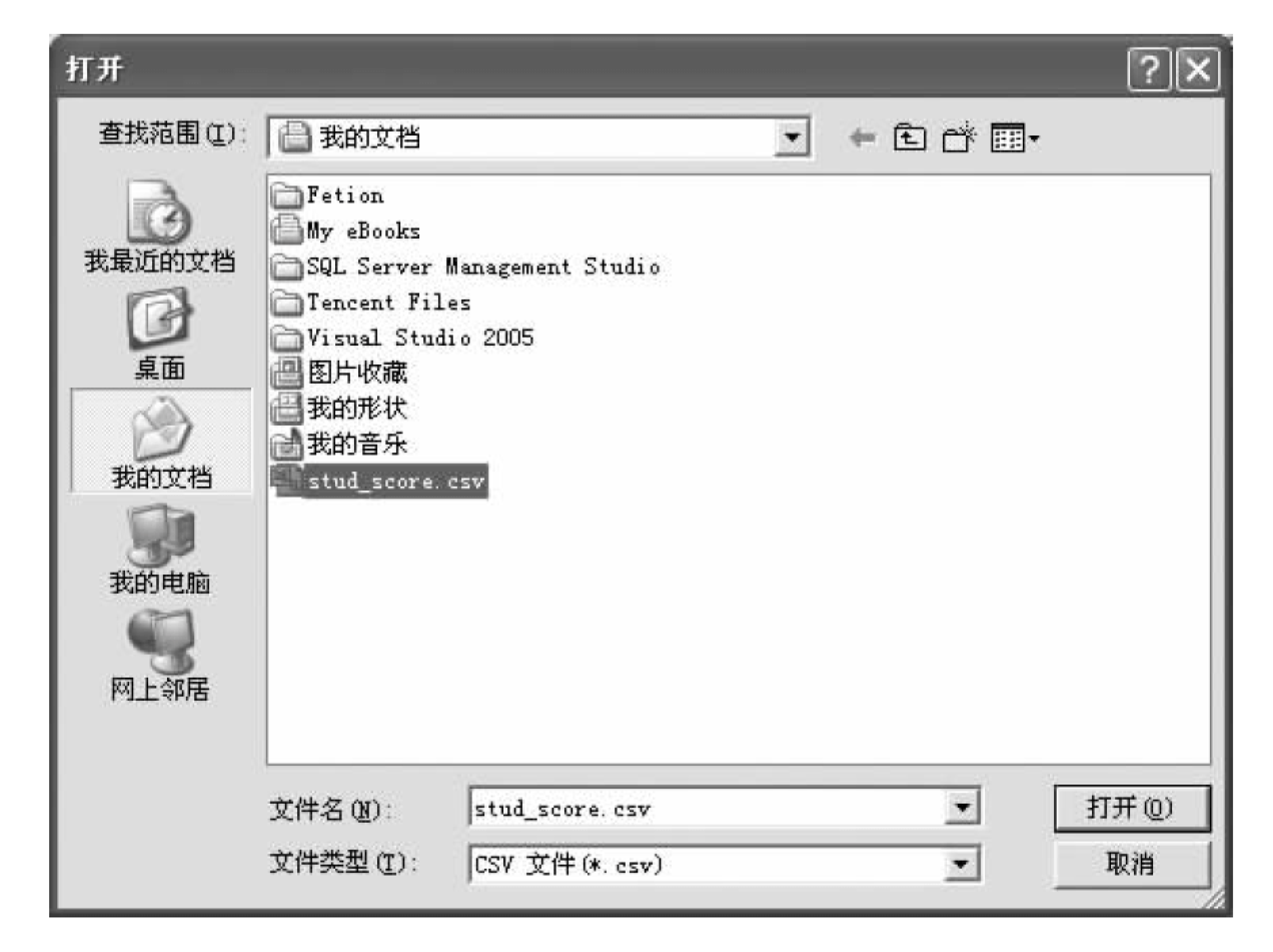


图 B-19 选择目标文件

(5) 单击“打开”按钮，回到导出向导，看到如图 B-20 所示的信息。由于在转换为 ARFF 文件时，Weka 必须从 CSV 文件的第一行读取属性名，否则就会把第一行的各属性值读成变量名，所以需要选中“在第一个数据行中显示列名称”复选框，保持对话框中其他选项为默认状态，单击“下一步”按钮。

(6) 在弹出的“指定表复制或者查询”页面中，选择“复制一个或多个表或视图的数据”单选按钮，并单击“下一步”按钮，如图 B-21 所示。如果要导出的数据不是来自于一个表，而是从一个或多个表中选择符合某些条件的数据，则选择“编写查询以指定要传输的数据”单选按钮。



图 B-20 进行格式设置



图 B-21 指定表复制或查询

(7) 在弹出的“配置平面文件目标”页面中，单击“预览”按钮，查看选定的数据表以及表中的数据，用以确认选择是否正确，如图 B-22 所示。

(8) 在弹出的“配置平面文件目标”页面中，单击“编辑转换”按钮，指定导出到目标文件时的操作，并单击“确定”按钮，如图 B-23 所示。此选项组有 3 个选项：“创建目标文件”、“删除目标文件中的行”和“向目标文件中追加行”单选按钮。如果之前导出的目标文件在磁盘上不存在，则此处默认选择“创建目标文件”选项，并且屏蔽另外两个选项。如果已经有该目标文件，这可以选择“删除目标文件中的行”单选按钮，用本次导出的数据覆盖原有的数



图 B-22　配置平面文件目标

据,或者指定"向目标文件中追加行",将本次导出的数据添加到目标文件已有数据的末尾。

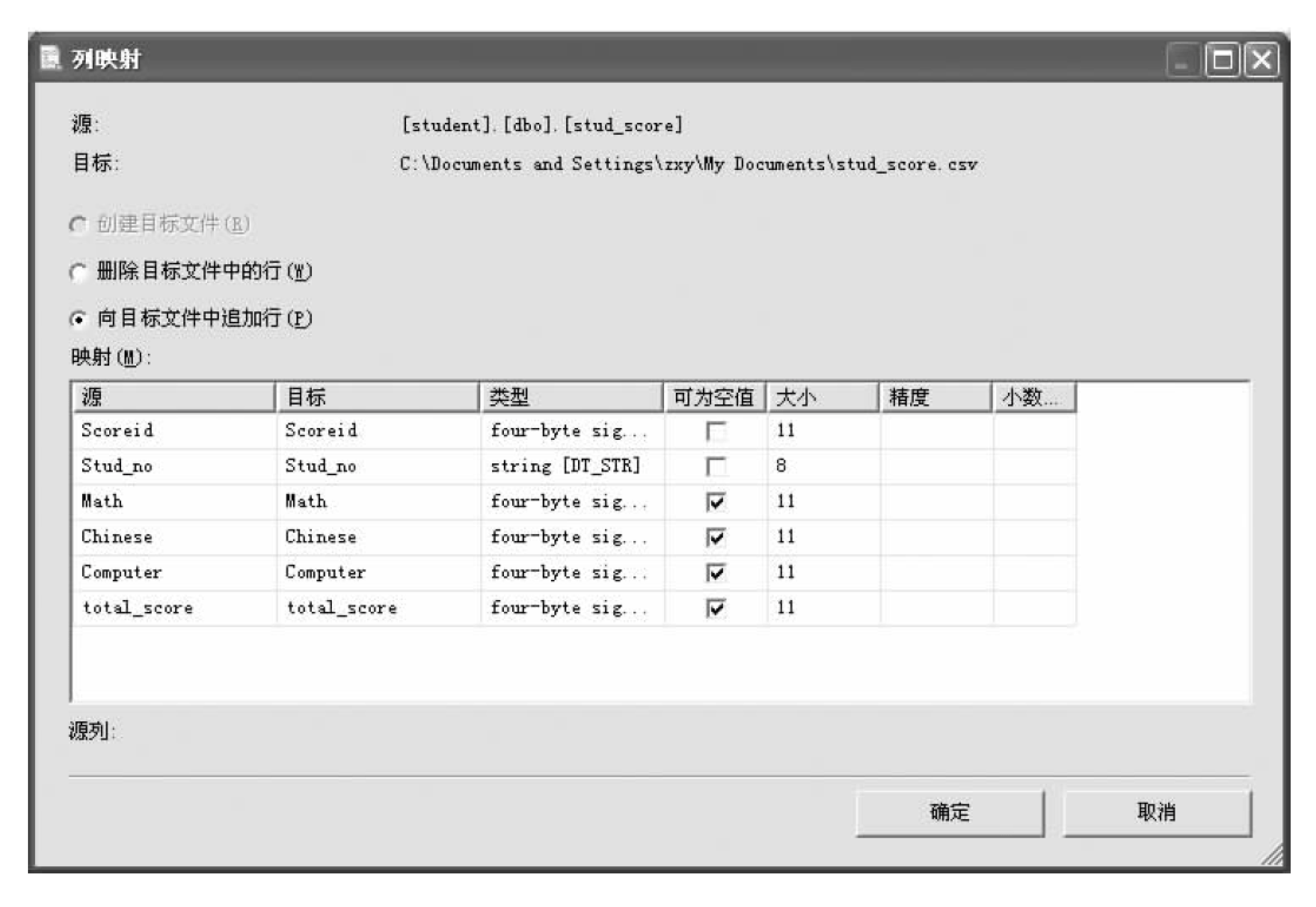


图 B-23　设置列映射

(9) 选中"立即执行"复选框,并单击"完成"按钮,执行导出,如图 B-24 所示。

(10) 导出成功显示如图 B-25 所示的对话框,至此,完成了将 SQL Server 2005 数据库文件导出成为 CSV 文件的过程。

(11) 用记事本打开 CSV 文件查看文件内容,如图 B-26 所示。

(12) 运行 Weka 的主程序,如图 B-27 所示。

(13) 单击进入 Simple CLI 模块,Simple CLI 模块是 Weka 提供的命令行功能界面,如图 B-28 所示。

SQL Server 导入和导出向导

保存并执行包

指示是否保存 SSIS 包。

☑ 立即执行(E)

☐ 保存 SSIS 包(S)

◉ SQL Server(Q)

○ 文件系统(F)

包保护级别(L):

使用用户密钥加密敏感数据

密码(P):

重新键入密码(R):

帮助(H)　< 上一步(B)　下一步(N) >　完成(F) >>|　取消

图 B-24　执行转换

SQL Server 导入和导出向导

执行成功

成功　12 总计　12 成功　0 错误　0 警告

详细信息(D):

操作	状态	消息
正在初始化数据流任务	成功	
正在初始化连接	成功	
正在设置 SQL 命令	成功	
正在设置源连接	成功	
正在设置目标连接	成功	
正在验证	成功	
准备执行	成功	
执行之前	成功	
正在执行	成功	
正在复制到 C:\Documents and Settings\zxy\My Do...	成功	已传输 6 行
执行之后	成功	
清除	成功	

停止(S)　报告(R)

关闭

图 B-25　执行成功界面

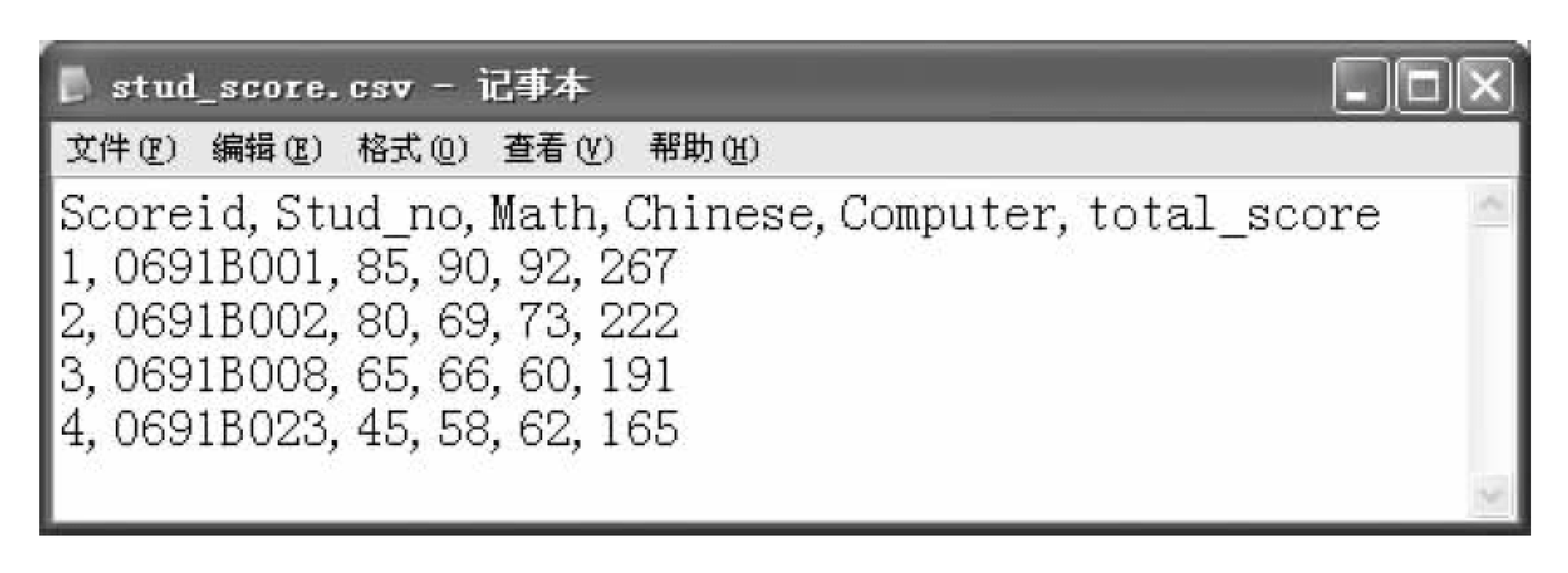


图 B-26 查看文件内容

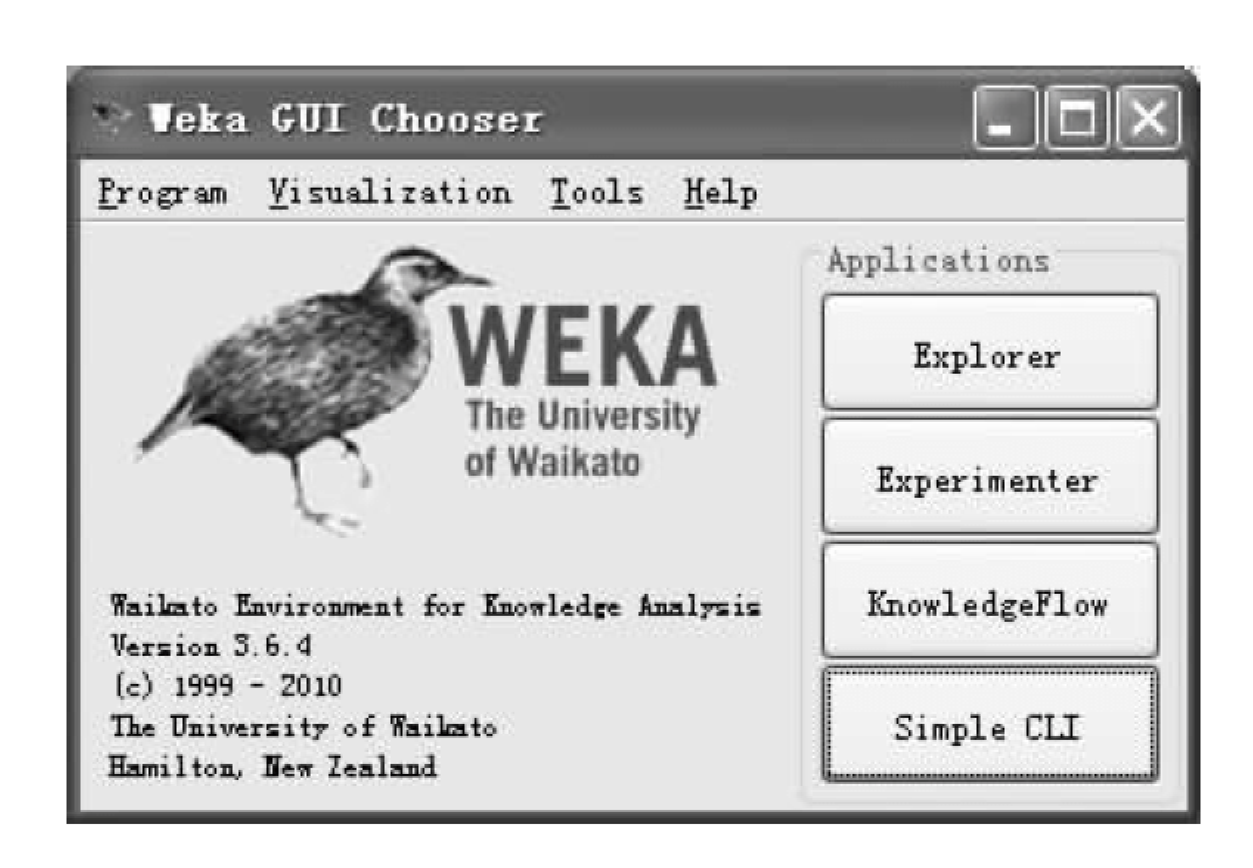


图 B-27 打开 Simple CLI 应用

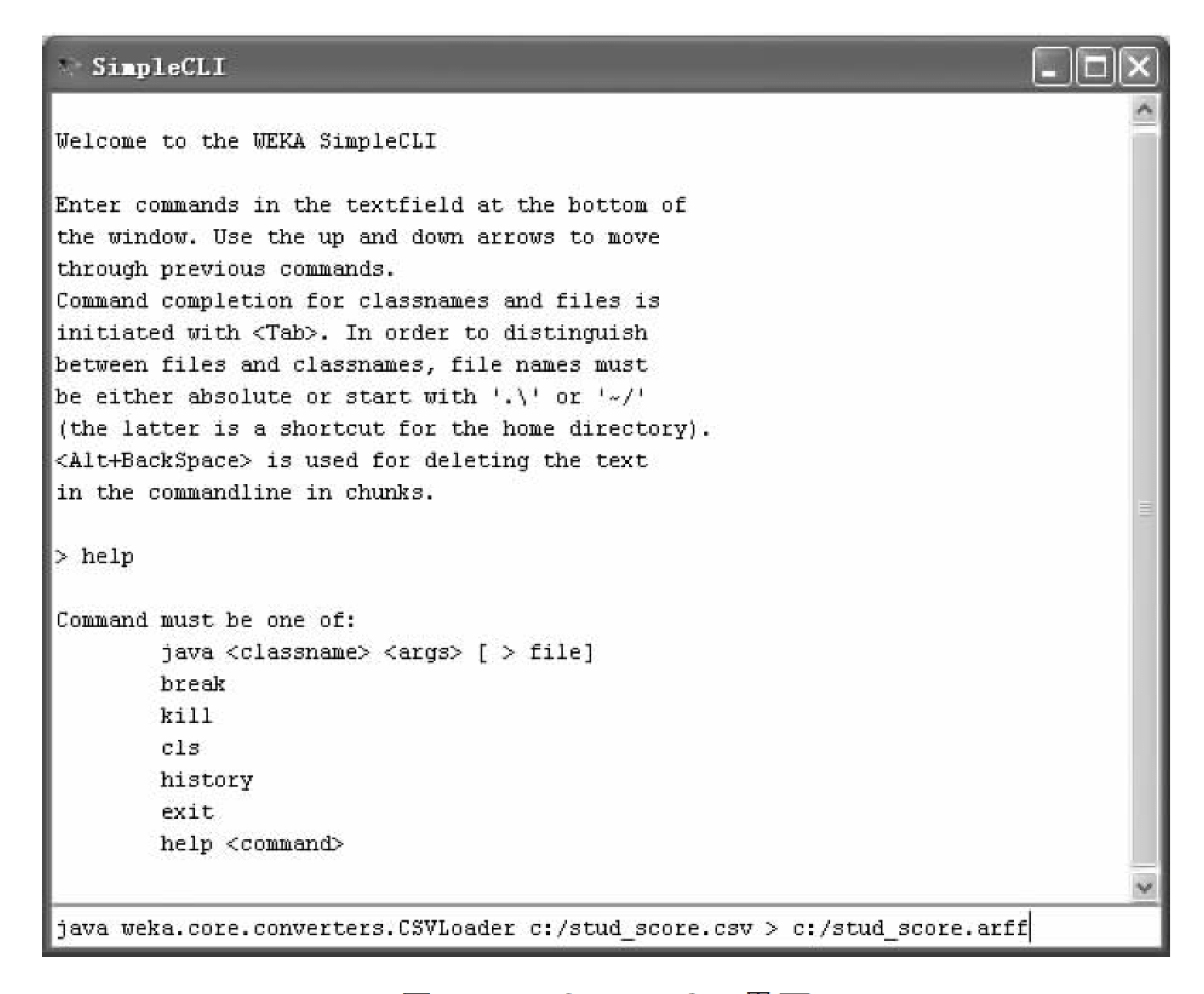


图 B-28 Simple CLI 界面

(14) 在新窗口的最下方有一行输入框，在这里可以输入命令。在此命令输入框中输入命令 java weka. core. converters. CSVLoader filename. csv > filename. arff 即可完成转换。其中 java weka. core. converters. CSVLoader 是 Weka 的一个命令，用来进行文件格式的转换，filename. csv 是待转换的 CSV 文件，filename. arff 是转换之后的 ARFF 文件。对于上面得到的 CSV 文件，假设放在了 C 盘根目录下，需要输入的命令是 java weka. core.

converters.CSVLoader c:/stud_score.csv >c:/ stud_score.arff，转换之后的文件名为 stud_score.arff。可以从 Simple CLI 窗口中看到转换的执行情况，如图 B-29 所示。

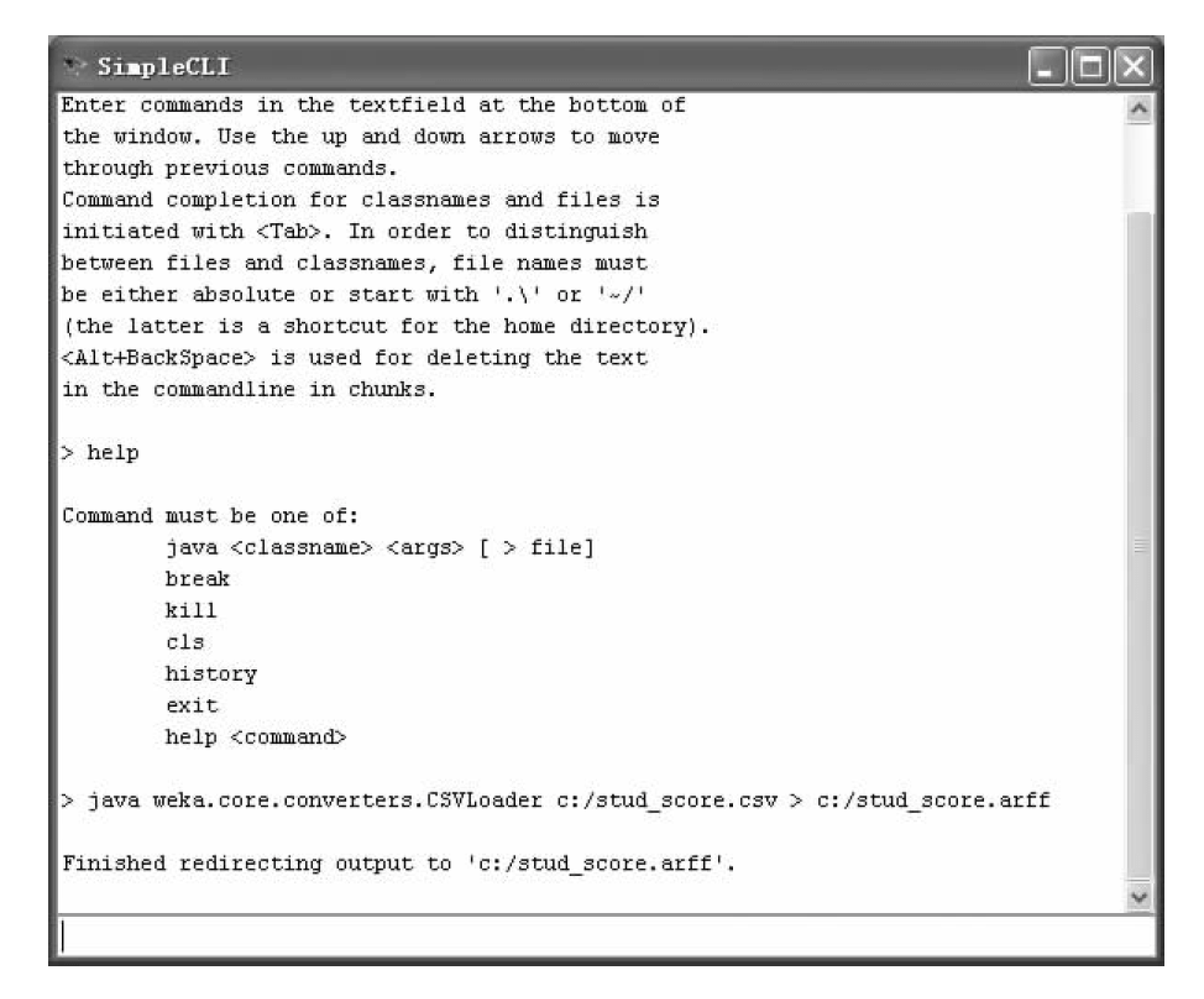


图 B-29　进行文件转换

（15）在 Weka 中还提供了 ArffViewer 模块，可以用它打开一个 CSV 文件进行浏览，然后另存为 ARFF 文件。从图 B-27 的主界面进入 Explorer 模块，单击上方的 Open files 按钮中打开 CSV 文件，如图 B-30 所示。然后点击右上方的 Save 按钮，将 CSV 文件另存为 ARFF 文件亦可。

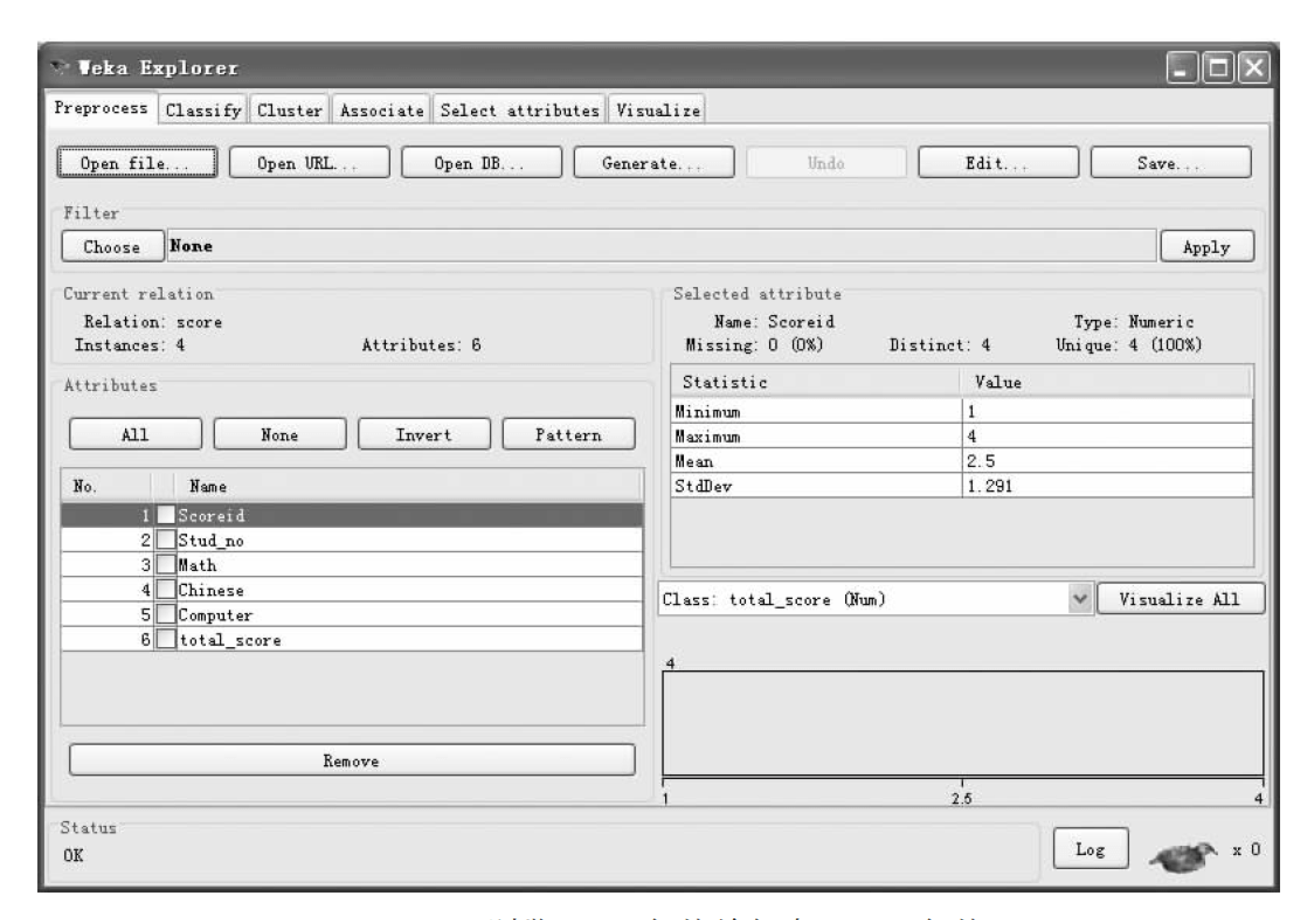


图 B-30　浏览 CSV 文件并保存 ARFF 文件

参考文献

[1] 张兴会. 数据仓库与数据挖掘技术[M]. 北京：清华大学出版社，2011.

[2] 姚志勇. SAS编程与数据挖掘商业案例[M]. 北京：机械工业出版社，2010.

[3] 王欣. SQL Server 2005数据挖掘实例分析[M]. 北京：水利水电出版社，2008.

[4] HAN J，KAMBER M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001.

[5] 张兴会. 基于递阶对角神经网络的失业预测研究[J]. 数量经济技术经济研究，2002，19(9)：114-117.

[6] 张兴会. 主成分分析法在神经网络经济预测中的应用[J]. 数量经济技术经济研究，2002，19(4)：22-125.

[7] 张兴会. 基于对角Elman神经网络的失业智能预测模型[J]. 南开大学学报，2002，35(2)：60-64.

[8] 张兴会. 基于神经网络模型的非线性多步预测学习控制器[J]. 控制与决策，2002(17)：820-823.

[9] 张兴会. 基于神经网络误差补偿的混沌系统广义预测控制[J]. 南开大学学报，2004，37(2)：89-92.

[10] 刘玲，张兴会，梁伟. 一种基于Aihara神经元的改进CBAM模型[J]. 天津工程师范学院学报，2009，19(2)：12-14.

[11] Du Shengzhi，Chen Zengqiang，Yuan Zhuzhi，et al. Sensitivity to noise in bi-directional associative memory(BAM) [J]. IEEE Trans. on NEURAL NETWORKS，2005，16(7)：887-898.

[12] 刘玲，张兴会. 基于神经网络的数据挖掘算法研究[J]. 计算机工程与应用，2008(9)：563-564.

[13] 王明春. 基于相对距离的改进粗k-means方法[J]. 计算机应用，2009(4)：1102-1105.

[14] 王明春，王正欧. 一种基于CHI值特征选取的粗糙集文本分类规则抽取方法[J]. 计算机应用，2005(5)：1026-1028.

[15] 王明春，王正欧. 基于粗糙集和遗传算法相结合的文本模糊聚类方法[J]. 电子信息学报，2005(4)：548-551.

[16] Zheng Xiaoyan，Sun Jizhou. Finding Frequent Item Sets from Sparse Matrix[J]. 2009 International Conference on Electronic Computer Technology.

[17] 郑晓艳，石连栓，孙济洲. 基于VOPP并行编程环境的最大频繁项集生成方法[J]. 计算机应用研究，2009(4).

[18] Tong Yongmu，Zheng Xiaoyan. An Algorithm to Construct FP-Tree on VODCA[C]. 2010 Third International Conference on Intelligent Computation Technology and Automation.